Pro gradu

Characterization and low-resolution structure studies of Transmembrane prolyl 4-hydroxylase

Aleksi Sutinen

University of Oulu

Faculty of biochemistry and molecular medicine

2018 Oulu

Table of Contents

Acknowledgements

I would like to express my gratitude to my supervisors Prof. Peppi Karppinen, Dr. Kristian Koski and Dr. Matti Myllykoski for their guidance and support. I thank Peppi for the overall guidance and making sure that, I have access to everything I need. I thank Matti for teaching the necessary skillset that a biochemist needs to conduct experiments. I thank Kristian for the mental support and the guidance on how to become a scientist. I feel privileged that I had this opportunity to work with all the people in Karppinen group, Anna, Joona, Riikka, Niina, Jenni, Raisa, Elitsa, Teemu, Tapio, Tanja, Eeva and Tuomas. I would also like to acknowledge the people in Dr. Rajaram Venkatesans group for the time we shared, Ramita, Abhi, Shiva, Sruthi, Subhadra, Gabriele, Ville, Pooja, Dihren, Tiila, Ed and Mikko. I also express gratitude to Prof. Rikkert Wierenga and Dr. Lari Lehtiö for the useful discussions.

Most importantly I am thankful for my partner Marja, who has supported me though all these years of studying.

- Aleksi Sutinen

2-OGDD	Fe (II)-/-2-oxoglutarate dependent dioxygenases
2-OG	2-oxoglutarate
DSBH	Double stranded β-helix
P4H	prolyl 4-hydroxylase
HIF	Hypoxia inducible factor
HIF-P4H	HIF prolyl 4-hydroxylase
FIH	Factor inhibiting HIF
ARNT	Aryl hydrocarbon receptor
bHLH	basic loop helix loop
PAS	Periodic circadian protein-Arnt (aryl hydrocarbon receptor)-Sim
CBP/p300	cAMP response element binding protein / histone acetyltransferase
NTS	nuclear translocation domain
C-TAD	C-terminal transactivation domain
N-TAD	N-terminal transactivation domain
ER	Endoplasmic reticulum
C-P4H	Collagen prolyl 4-hydrohylase
P4HTM	Transmembrane prolyl 4-hydroxylase
Cr-P4H	*Chlamydomonas reinhardtii* prolyl 4-hydroxylase
4Hyp	4-hydroxy proline
pVHL	Von Hippel Lindau tumor suppressor protein
E3	ubiquitin ligase
HRE	Hypoxia responsive element
EPO	Erythropoietin
ECM	Extracellular matrix
ANK	Ankyrin repeat
EF-hand	Helix E/Helix F hand
PDB	Protein databank
PSB	Peptide substrate binding domain
TauD	Taurine hydroxylase
CD	circular dichroism
SEC	size exclusion chromatography
MALS	multiangle light scattering
QELS	quasi elastic light scattering
d.p.m	discintillation per minute
MX	macromolecular crystallography
SAXS	Small-angel X-ray crystallography
Rg	radius of gyration
Dmax	maximum particle distance
P(r)	Distance distribution function
MW	molecular weight
N-linked	N-acetyl glucosamininated residue

1. INTRODUCTION

Oxygen is the most vital component of life. From the lowest trophic level to highest, almost all species require oxygen. Oxidative phosphorylation is the essential machinery that utilizes energy from oxygen, available for cells in the form of adenosine triphosphate. However, there are coping mechanisms to protect cells when oxygen supply is decreased — physiological state where the partial pressure in the cells decreases below 6% is called hypoxia. During hypoxia, cells initiate a hypoxic response, which is a complex system that utilizes multiple anaerobic signal transduction pathways in order to cope with the oxygen loss. The most crucial role during hypoxia is the stabilization of hypoxia-inducible factor (HIF). HIFs are transcription factors that induce the expression over 300 target genes, related to glycolysis, vascular perfusion, erythropoiesis, cell proliferation, and growth. The regulation of HIFs and oxygen sensing are, regulated by HIF prolyl 4-hydroxylases. These enzymes regulate the stability of HIF by hydroxylation of proline residues in the HIF oxygen-dependent degradation domain. Under normoxia, HIF is transported to ubiquitin targeted proteasomal degradation.

Prolyl 4-hydroxylases belong to the iron- and 2-oxoglutarate dependent dioxygenases (here on abbreviated as 2-OGDD). Enzymes belonging to this group utilizes iron with a redox state II in coordination with 2-oxoglutaric acid (abbreviated 2-OG) and molecular oxygen. The coordination of the active site is formulated by conserved double-stranded β-helix fold, also referred to as the ’jelly-roll fold’ or β-sandwich. The interaction of 2-OGDD substrate molecules is occurring through iron-coordination with 2-histidine-1-carboxylate (H-X-D-H)motifs. Enzymes catalyze the addition of hydroxyl group to 4R,2S proline residues in their substrates. This project studies the structural and protein dynamics of the ER associating transmembrane prolyl 4-hydroxylase (abbreviated P4H-TM). It is the second prolyl 4-hydroxylase localized in the endoplasmic reticulum (ER), and the only member associating with the membrane. To this date, the structural and functional features of P4HTM has remained unclear. The enzyme has a high sequence homology towards human collagen prolyl 4-hydroxylases and HIF Prolyl-4-hydroxylases, and acts on the similar substrate, but may have additional physiological roles.

2. REVIEW OF LITERATURE

2.1 Fe (II)-/-2-oxoglutarate dependent dioxygenases activates oxygen in an array of biochemical reactions

Transition metals such as manganese, copper, and iron are effective co-factors in biocatalysis when bound to active site residues and other co-factor molecules, water, molecular oxygen, and nitrogen compounds (Karlin. 1993). Metalloenzymes coordinate transition metal cations into the active site and is a well-understood strategy to catalyze a variety of biochemical reactions (McCall *et al.* 2000, Vallee & Williams. 1968). Molecular oxygen itself rarely can take part in biochemical reactions spontaneously in cellular systems, when reactive oxygen species (ROS) are not taking to account. Fe(II)-/- 2-oxoglutarate dependent dioxygenases (hence abbreviated as 2-OGDDs) utilize 2-OG, non-heme ferrous iron (redox state Fe(II)) and molecular oxygen to oxidize substrates including protein substrates and other biomolecules such as genomic DNA and RNA (Hausinger. 2004) Oxygen is a very electronegative element, making it a powerful nucleophile for biocatalytic reactions when coordinated with a complex. These enzymes are described as dioxygenases because of both electron pairs from the oxygen heteroatom is utilized to oxidize the substrate and 2-OG co-substrate molecules. The other electron pair ens up always into the succinate when 2-OG decarboxylates. The second lone pair of oxygen is used to the oxidization of the actual primary substrate.

2.1.1 Double-stranded ß-helix fold

2-OGDDs are metalloenzymes which coordinate iron with active site residue in a hydrophobic core, which refers to as the double-stranded β-helix fold (Hence abbreviated as DSBH-fold) (Loenarz & Schofield. 2011). The fold consists of eight antiparallel β-sheets which form distorted two-handed barrel shape helical-like barrel shape. The sheet ϕ- and ψ-angles generally satisfies the Ramachandran angles, although the sheets can twist and twine, thus term distorted. The β-sheets are dived into minor (β1-, -2,-7,-4) and major grooves phasing each other (β1,-8,-6,-5). The β-sheets belonging to the core fold are often listed separately from other β-sheets with a Roman number, e.g., βI-VIII. (Clifton *et al.* 2006, Islam *et al.* 2018, McDonough *et al.* 2010)

The helical barrel is more open towards to the catalytic metal binding site and allowing substrate access the core. In most 2-OGDD structures the major groove βI- and βVIII-sheets are located at the more open end, leading the substrate binding. The same type of fold is seen on plant cup in-fold proteins and Jumonji C chromatin (JmjC) associated proteins. Usually, the major sheets are more extended respect to the nation-DSBH β-sheets, which formulates the core shape. The minor β-sheets usually share the size of non-DSBH sheets and can be even shorter. However, there are several exceptions observed within the subfamilies (Schofield & Zhang. 1999). Another distinctive feature is the linking loops, especially between βII-βIII-sheets and often right before the β1 or between βI-βII. The βII-βIII loop may appear as a short flexible random coil or contain much ordered larger even domain-like structures (L. Hu *et al.* 2013, Kato *et al.* 2011, McDonough *et al.* 2006, Zhang *et al.* 2004). Also, the C-terminal part after the DSBH-fold also has been shown to undergo a conformational change during substrate binding (Chowdhury *et al.* 2009). These regions vary the most in the superfamily, and often explaining the substrate specificity and vary within the subfamilies (Aik *et al.* 2012). Another differentiating feature among the subfamilies is the N-terminal part before the core fold. The core fold is supported at least two N-terminal helices, which supports the fold since the core is hydrophobic. The DSBH-folding features may resolve from insertions and turns between the sheets, especially in the sheets that form the 'pocket' of the core (Hewitson *et al.* 2005, Koski *et al.* 2007, McDonough *et al.* 2006). The volume of the which the DSBH-fold may take could explain the substrate selectivity.

2.1.2 Substrate types

Since the DBSH-fold allows broad substrate selectivity, 2-OGDDs have a broad group of substrates. The number of these enzymes in animals (over 80) and especially in plants, explains the involvement of 2-OGDDs in many types of reactions with many different substrates (Clifton *et al.* 2006). All 2-OGDD oxidization reactions require Fe^{2+} and molecular oxygen. However, 2-OG itself is not always included in the catalysis (Lundberg *et al.* 2008). The variety of substrates includes small molecules such as antibiotic compounds, fatty acid moieties, Nucleic acids, and proteins. The reaction types catalyzed by 2-OGDDs and related enzymes include hydroxylation, demethylation, desaturation, epimerization, epoxidation and alkyl additions to cyclic substituents.

The earliest findings of a 2-OGDD catalyzed reaction were identified as hydroxylation of protein substrates with structural roles. The collagen synthesis was shown to involve proline hydroxylation catalyzed by prolyl-hydroxylases (C-P4H) (Bhatnagar *et al.* 1967, Goldberg & Green. 1968, Hutton *et al.* 1966, Hutton & Udenfriend. 1966, Kivirikko & Prockop. 1967). Decades later 2-OGDD enzyme was found to have a role in oxygen sensing. The key regulator of animal hypoxia response regulator HIF1α stability was linked to proline hydroxylation (HIF-P4Hs) and with a separate mechanism to asparaginyl hydroxylation (FIH) (Bruick & McKnight. 2001, Dann *et al.* 2002, Semenza & Wang. 1992). Protein substrates apart from previous roles, hydroxylation is also seen in ribosomal-, RNA splicing-, and other nuclear proteins. These proteins often contain the Jumonji domain (JMJD). In humans, ribosomal hydroxylations are catalyzed by MINA53 and NO66, which hydroxylate 3S-histidyl residues (Chowdhury *et al.* 2014). The second type of modification of involved in gene expression is the N-lysine demethylation of histone proteins, namely as KDMs which contain Jumonji domain type C (JmjC) (Agger *et al.* 2007, Chen *et al.* 2007, Couture *et al.* 2007). Some hydroxylases have parallel roles. FIH, which catalyzes the hydroxylation of Asparagine residue in HIF1α c-terminal transactivation domain (CTAD), also acts on ankyrin repeats found in Notch-receptors. The Notch hydroxylation links FIH to a cell signal transduction system apart from HIF (Coleman *et al.* 2007).

Many 2-OGDDs are involved in gene regulation by acting on the DNA and RNA molecules itself. The most well-known mechanism was identified from bacteria and the methylation of adenine and cytosine bases, preventing alkylation damage. The reaction is catalyzed AlkB demethylases which were identified in *E.coli* and had many homologs in *eukaryotes* including humans (Kurowski *et al.* 2003, Trewick *et al.* 2002). In some cases, the substituent, instead of the hydroxyl group can be a halogen such as chlorine or bromine. These reactions are typical in antibacterial-, antifungal-, antiviral- and anti-inflammatory synthesis such as threonine chlorination catalyzed *Pseudomonas syringae* B301D (SyrB2). The reaction is part of much crosslinking halogenation, also known as the Syr1/Syr2 system in the synthesis of syringomycin E. (Blasiak & Drennan. 2009, Vaillancourt *et al.* 2005). As discussed in section 2.1.1 the abundance of 2-OGDDs is more imminent in plants than animals. An example of lipid metabolism is the Phytanoyl-CoA hydroxylase, which is a 2-OGDD involved in converting fatty acid phytanoyl-CoA into pristanic acid(Croes *et al.* 1997). Pristanic acid is part of β-oxidization. In plants, 2-ODDS are involved in flavonoid and gibberellin synthesis, generation

of herbicide prospecting compounds and other plant-specific products. From here on hydroxylases, and especially prolyl 4-hydroxylases are discussed in detail.

2.2 The chemistry of 2-OG dependent dioxygenases

Carbon sources can be utilized and activated many ways in cellular processes. The typical strategy of oxygen activation of metalloenzymes consists binding molecular oxygen, to produce ferryl intermediates (Fe (IV) –oxo complexes), either to acts as an electrophile or a hydrogen donor (Kivirikko & Prockop. 1967, Pau *et al.* 2007). These properties are critical in biocatalysis. The catalytic properties of 2-OGDDs were first identified from collagens were the dioxygenase activity demonstrated to hydroxylate collagen chains (Kivirikko & Prockop. 1972, Majamaa *et al.* 1984). The first primitive reaction mechanism for the hydroxylation mechanism was described, by the work done with taurine hydroxylase (TauD)(Hanauske-Abel & Günzler. 1982). Since then, the proposal has held surprising accuracy supported by kinetic and thermodynamic experiments.

The critical aspect regarding efficient catalysis is the spin state of the iron ion and the co-substrates. The molecular oxygen molecule incorporates a triplet spin state in the O_2-molecule. Each oxygen atom possesses two lone pairs and one electron in the S_1 atomic orbital. Two oxygen molecules form molecular orbital, which constructs from the lower energy 2S atomic orbitals and two stable π2p-molecular orbitals. The remaining two electrons occupy the antibonding π2p*-molecular orbitals with opposing spins (Weller inorganic chemistry 6ed). The organic co-substrate 2-OG possesses singlet spin state (Hernández-Ortega *et al.* 2015, Weller *et al.* 2014). The orbital overlap and the repulsion is acceptable making the uncatalyzed reaction thermodynamically favorable. However, the kinetic rate of the uncatalyzed reaction is slow. Any other spin states possible in either molecule are not possible, and the reduction potential of the O_2-molecule is low and unfavorable, resulting in a slow reaction. The free energy transition state barrier is lowered with the construction of transition-metal-ion complex (Fersht. 1999, Wolfenden. 2014).

Through the redox process via the Fe-intermediate states, the oxygen molecule activates, and the 2-OG is fully unsaturated, resulting in a variety of different products. In order to drive the redox potential, the system requires electrons, which are provided by the organic molecule, in this case, 2-OG (Costas *et al.* 2004, Shan & Que. 2005). In most 2-OGDD enzymes as well as

other metalloenzymes, the non-heme iron is coordinated with two histidines and carboxylate groups usually from aspartate or glutamate, together with O_2. The binding position of the Fe^{2+} -ion is often referred to as the 2-His-carboxylate facial triad, due to the residue orientation (Islam *et al.* 2018). The keto-group faces either Asp/Glu-residues and the carboxylate histidine. Two water molecules take part in the coordination into Fe^{2+}, when 2-OG is not bound. The solvating two water molecules are displaced, when the co-substrate is bound to the complex. The primary substrate displaces the last water molecule coordinating Fe^{2+}. In the initial step of the reaction, iron is in 6-coordination with a redox state Fe(II).

As a result of the iron (Fe^{2+}) coordination in DSBH-core, and the binding of 2-OG and the oxygen addition causes more open conformation for the primary substrate binding. The fold could have a key role in the substrate specificity (Aik *et al.* 2012). The most well studied and understood mechanism of 2-OGDDs involves hydroxylation of the substrate (Borowski *et al.* 2004). The reaction mechanism of substrate hydroxylation resembles six steps, including three intermediate states (**Fig.1**). The electron transfer from the metal ion to oxygen causes a change in the redox state. The O_2 binding yields a transient Fe^{3+}-superoxo intermediate. The distal oxygen atom possesses a negative partial charge. The delocalization induces π2p-electron from the distal oxygen and attacking the C2 carbon atom of 2-OG carboxyl group, forming a peroxohemiketal bicyclic intermediate, where 2-OG decarboxylates, and CO_2 is released as a byproduct. As a result of the decarboxylation and the delocalization of the charge, a ferryl(IV)-complex forms which possesses a high S2-spin state which was shown applying Mössbauer spectroscopy (Hoffart *et al.* 2006). The ferryl state is the first observable intermediate state with a second-order rate 1.5 x 10^5 M^{-1} s^{-1}, decomposition at first order rate 13 s^{-1}, according to stopped flow UV-VIS spectroscopy (Krebs *et al.* 2007). These findings show that the intermediates are extremely transient and the reaction rates are fast when oxygen activates with the metal complex. The ferryl(IV)-complex carries enough redox potential to abstract the hydrogen from the primary substrate yielding a substrate radical. The metal complex takes Fe(III)-OH superoxo state, which is the second observable intermediate state (Price *et al.* 2005).

In the last step of the reaction, a rebound effect occurs, meaning the exchange of hydrogen atom, between the substrate radical and the Fe(III)-OH adduct. The rebound results in unsaturation of the peroxohemiketal bicyclic intermediate to succinate. Fe(III)-superoxo intermediate reduces back to the initial Fe(II)-state. The product is hydroxylated and dissociated

from the complex. The retention of Fe(III) to Fe(II) is the third observable state. However, there is still no evidence of the first Fe(III)-superoxo intermediate or the Fe(III)-OH species (Tarhonskaya *et al.* 2014b). Cryogenic Raman spectroscopy experiments have suggested an alternative dissociation path of the Fe(III)-OH species involving deprotonation of the Fe-OH from nearby oxo-groups. This non-observable step may cause the formation of possible Fe-alkoxide species, before the complex dissociation (Grzyska *et al.* 2010). The use of metal-complex to activate oxygen has shown to be a powerful approach to catalyze multiple biochemical reactions across the phylum, which explains the vast number of 2-OGDDs in nature (Martinez & Hausinger. 2015).

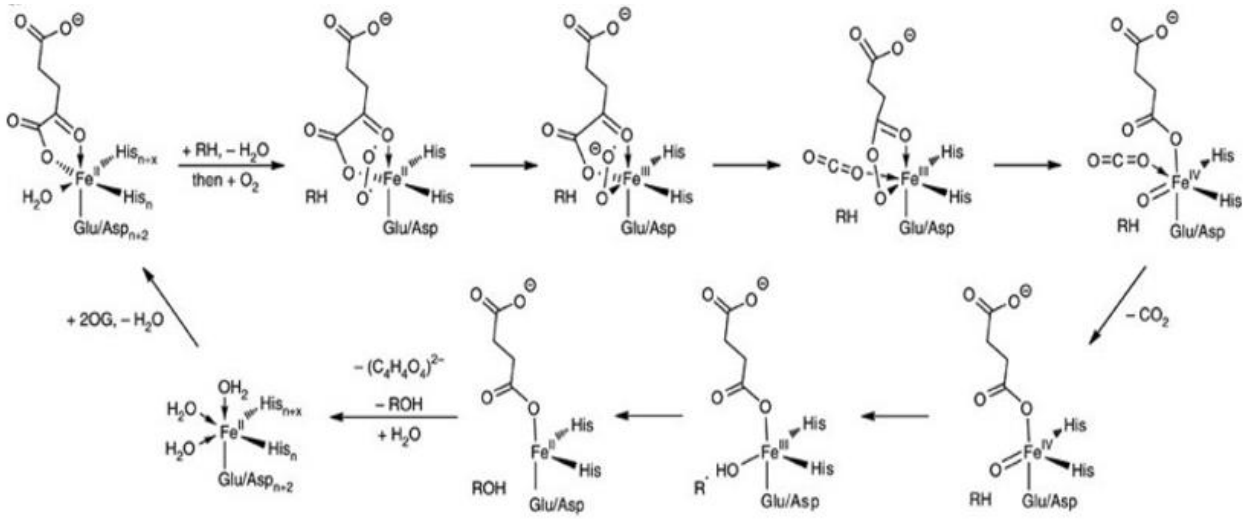


Fig.1 The reaction mechanism cycle of 2-OGDD enzymes. Fe^{2+} is bound to the 2-his-1-carboxylate facial triad. 2-oxoglutarate displaces solvent water and binds to the iron with keto-carboxyl group orientation facing toward Asp/Glu. Molecular oxygen attacks 2-OG C2 carbon yielding the Fe(III)-superoxo and peroxohemiketal intermediates. The charge delocalizes yielding Ferryl(IV)-complex with the high spin state. The primary substrate is in the vicinity of this highly polar complex, and hydrogen is abstracted from the substrate. The iron reduces to Fe(III)-OH species though vigorous rebound effect the substrate exits the binding pocket with the hydroxyl group and the byproduct as unsaturated to succinate. The figure was adopted from the publication (Tarhonskaya *et al.* 2014a), with the permission of the publisher.

2.3 Cells respond to hypoxia

Hypoxia response is an adaptation system of the cell to secure energy flux in hypoxic conditions for the cells. From a physiological aspect, hypoxia causes variable responses. Important tissues such as brains, acutely respond to hypoxia. These events are initiated in the nervous system, where the ionic strength shift of Ca^{2+}-/K^{+} ion pumps leads to depolarization of nerve cells, leading to peripheral vasodilation, pulmonary vasoconstriction, and a decrease of plasma volume. Visible symptoms of acute hypoxia are an increase in heart rate and breathing frequency. The increase of oxygen perfusion with previous events is an example of a fast response to hypoxia (Michiels. 2004). In the molecular level, anaerobic glycolytic pathways take over when the oxygen supply is not sufficient to maintain oxidative phosphorylation and ATP threshold. The most important response pathways are called the hypoxia-inducible factor-pathway (HIF-pathway) (Semenza. 2001).

HIF-pathway induces expression of genes which increase angiogenesis, erythropoiesis, cell growth, and proliferation. The primary regulatory element under hypoxia is the hypoxia-inducible factor HIF. It is a heterodimeric transcription factor. Some examples of the over 300 human HIF1α target genes are Vascular endothelial growth factor (VEGF), erythropoietin (EPO mainly HIF2α), platelet-derived growth factor (PDGF), Glucose transporter-1 (GLUT-1) which are upregulated under hypoxia (**Fig.2**) (Gordan & Simon. 2007, C. J. Hu *et al.* 2003, Kaelin & Ratcliffe. 2008, Tian *et al.* 1997). The hypoxia response pathway is present in most metazoans and highly conserved between flies, worms, and mammals. All animals produce some components in the HIF-pathway (Bishop & Ratcliffe. 2014). There are also integrated pathways along with HIF that are identified responding to hypoxia. In normoxic conditions, HIFα is hydroxylated by HIF-P4Hs leading to Von-Hippel Lindau tumor suppression protein (pVHL) mediated ubiquitous proteasomal degradation.

First proceeding was made when a Hypoxia-responsive element (HRE) was identified. It was first identified as a flanking region of the human Erythropoietin (EPO)-gene and showed to upregulate under hypoxia (Beck *et al.* 1991). The presence of HRE leads to the discovery of many other co-activators affecting the EPO-transcription activity. The most important discovery was made by Semenza and Wang (1992) when they reported a novel EPO regulator. The novel regulator was named as HIF, which associated with HRE. The HIF1α forms a

heterodimer with a co-expressed HIF1β referred to as aryl hydrocarbon translocator (ARNT). The HIFαβ-dimer binds to HRE together with two co-transcription factor complexes. Histone acetyltransferase P300 (P300)/cAMP response element binding protein (CBP) co-transcription factor complex (Liu *et al.* 2008) and steroid receptor activator 1 (SRC-1) (Onate *et al.* 1998). It should be noted that the hypoxia initiates multiple integrating pathways, which are not related to the HIF-pathway. Epigenetic changes can also mediate HIF stability.

Additionally, some transcription factors, G-protein coupled receptors which are activated in hypoxia and are not part of HIF-pathway and adopt a similar strategy. Some examples are NF-κB, which is involved in cytokine synthesis and other cellular stress. NF-κB is regulated with inhibitory kinase IκB, which contains the LXXLAP HIF target sequence (Culver *et al.* 2010). Adrenergic receptor subtype β_2 (β_2AR) has shown to be hydroxylated in hypoxia and also mediated by pVHL-E3 ubiquitous proteasomal degradation (Xie *et al.* 2009). Animals also have survival systems with an entirely different mechanism than HIF-pathway (such as SIRT1 in humans) which are common genes related to cell survival (Guarente 2011). However, the role of previous genes in hypoxia and the crosstalk between HIF-pathway is still unclear, and interaction between HIF-P4H have not been observed (Gorres & Raines. 2010).

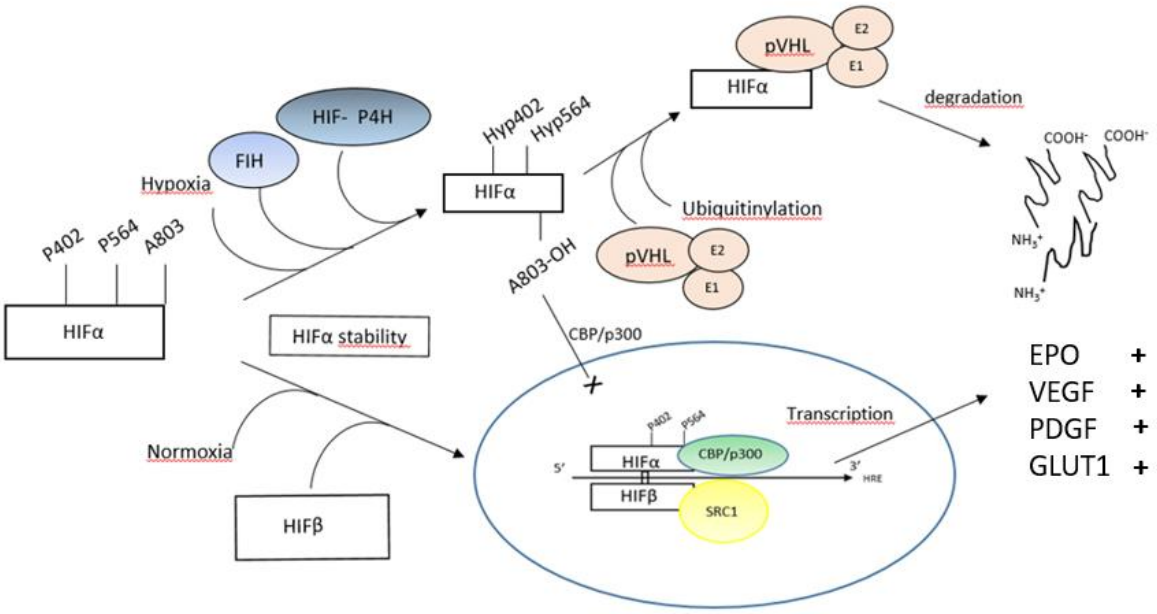


Fig.2 HIF-pathway. HIF1α is expressed continuously into the cytoplasm with a short lifetime. Under hypoxia, HIF1α stabilizes and dimerizes with HIFβ, and the complex is transported to the nucleus. The HIF-dimer associates to HRE with CBP/p300-, and SRC1 inducing target genes, leading to hypoxia response. In normoxic conditions, HIF prolyl 4-hydroxylases

hydroxylate proline residues in oxygen-dependent degradation domain (ODDD), leading to pVHL-linked ubiquitous protein degradation. Additionally, factor inhibiting HIF, hydroxylates asparagine residue in C-terminal translocation domain (CTAD), preventing CBP/p300 interaction, and suppressing HRE mediated gene expression.

2.3.1 Hypoxia-inducible factor

In hypoxia, the prolyl-4-hydroxylase activity is inhibited, and the HIF1α stabilizes and forms a dimer with HIFβ. The HIF-dimer binds to the HRE promotor region, with the p300/CBP-module (cAMP response element binding protein). The p300/CBP-complex contains histone acetyltransferase activity which is a characteristic feature of among transcription factor.

The HIF1α and HIFβ belong to the basic helix-loop-helix (bHLH) proteins that are common transcription factors, often coupled with a dimerization PAS-domain (Pellequer *et al.* 1998). First identified in *Drosophila* in the periodic clock proteins PER, together with ARNT and SIM-domains, the main purpose of the module is protein-protein interactions and ligand interactions (Pellequer *et al.* 1998). The DNA binding of bHLH proteins mostly binds 5'-TACGTG-3 segments in the gene, which is also present in the HRE. The HIFα amino-terminal part is comprised of two ARNT-PAS-SIM domains PAS A- and PAS B-subdomains, which are highly conserved in all HIF paralogues (G. L. Wang *et al.* 1995). The heterodimer of HIFα and HIFβ is formed with antiparallel β-sheet contact between the PAS B subdomains. The nuclear localization signal (NLS) in both ends is required to transport HIFα from the cytosol into the nucleus (Jiang *et al.* 1996). The proline hydroxylation of HIFα occurs in proline residues in the ODDD, LXXLAP motifs. In HIF-P4H2 these prolines are located in P313 and P374 (McDonough *et al.* 2006)

There are three different HIFα isoforms found in a human which all have different α-subunits (Semenza. 2000). HIF1α being the most prominent and firstly characterized pleiotropic transcription factor active in acute hypoxia. The targeted genes of HIF1α encodes short-term induction of glycolytic pathway such as pyruvate kinase in glycolysis. HIF1α is found to be expressed in all cells types, with a short lifetime (Moroz *et al.* 2009, G. L. Wang *et al.* 1995). HIF2α is expressed in multiple cell types but has been identified to upregulate genes that are more involved in long-term hypoxia response, such as EPO induction (C. J. Hu *et al.* 2003, Ratcliffe. 2007) However, the role of least studied isoform HIF3α is still unclear. However, it

contributes to the HIF-signaling pathway to some extent but may have some other unknown roles. Other isoforms HIF2α and HIF3α form dimers with HIFαβ-dimers, where the β-domain is identical as with HIF1α (Kaelin & Ratcliffe. 2008).

Before a comprehensive understanding of HIFα destruction, it was noticed that metal cations such as Co^{2+}, Ni^{2+} and Mn^{2+}, and chelating compounds such as EDTA upregulated HIF1α initiating a hypoxic response. Hypoxic sensing of the cell is a result of the stabilization of HIF1α. Under normoxic conditions, HIF-P4Hs can hydroxylate the proline residues in the C-terminal and N-terminal degradation domains of HIFα. The hydroxylated HIFα then binds to the von-Hippel Lindau protein, which acts as a targeting protein for Ubiquitin E3 ligase system, leading to HIFα degradation. Active form, pVHL consist of α- and β-domain with a flexible linker. The ubiquitin targeting complex also consists Elongin B and – C, which takes part in the binding of HIFα (Ivan *et al.* 2002, Min *et al.* 2002, Ohh *et al.* 1999). Other components of this complex include Elongin B, -C, Cullin box2 (Cul2), and ring box protein 1 (Rbx1) (L. E. Huang *et al.* 1998, Salceda & Caro. 1997).

2.3.2 Extracellular matrix and collagen

The extracellular matrix (ECM) frames the individual cells increasing cohesion and integrating them into tissues. The body of ECM is constructed from robust elongated and mostly large fibrillar proteins, such as collagens, proteoglycans, laminins, and fibronectins. Each of these proteins is present in multiple types, and each family carries many different tasks. ECM is an active, mobile cellular compartment. The joined structure of these abundant proteins forms a network, which provides mobility, flexibility, and structure to the tissue. The biophysical and biochemical properties of ECM mostly depict the tissue-specific function. The network system forms can form physical links between cell, such as desmosomes or tight junctions (Mouw *et al.* 2014).

Collagens are trimeric fibrous proteins that from long triple-helical polypeptide chains (Gordon & Hahn. 2010). It is the most abundant protein in vertebrates by mass. The functional repeating units of collagens consist of $[G\text{-}X\text{-}Y]_n$-units, which can extend to over 1000 residues long (Myllyharju & Kivirikko. 2004). The abundant appearance of glycine allows the collagen molecule to be flexible and elastic, due to the small size and low steric hindrance of the

molecule. The other two residues in the X and Y positions are often proline or hydroxyproline residues.

In some cases, the proline may be replaced by any other residue, but only lysine residues are hydroxylated in addition to proline residues (Myllyla *et al.* 2007a). Hydroxyproline is a modified proline residue of proline, containing an additional hydroxyl group in either C3 or C4 carbon, obtaining C-endo or C-exo conformation. So far in humans, there are 28 different types of collagens, namely separated with roman numbers. They vary both by their function and structure. They may have different oligomerization states, such as collagen type II which forms homotrimers, or types I and V which appear as heterotrimers. Collagens are categorized structurally into fibrillar collagens and non-fibrillar collagens. The most distinct structural difference is that non-fibrillar collagens contain at least one disordered motif in the structure. Types of non-fibrillar collagens consist associated fibril collagens with interrupted triple helices- (FACITs), short chain-, basement membrane-, multiplexing-, membrane-associated collagens with interrupted triple helices- (MACIT) and other types of collagens. Defects in these collagen structures often lead the cause of connective tissue disorders and diseases. Even tough collagens are the main components of ECM; they can appear as intracellular forms (Brodsky & Persikov. 2005, Myllyharju & Kivirikko. 2001, Ortega & Werb. 2002).

The triple peptide unit assembly dictates the structure and function of the particulate collagen type. Due to the large size of collagen molecules, the chains translated into the ER as unmatured pro-collagen single chains. Protocollagens are primary forms of collagens which require post-translationally hydroxylation. The post-translational modification includes hydroxylation of proline residues and lysine residues. These reactions are catalyzed by collagen proline-4-hydroxylases (C-P4H), prolyl-3-hydroxylases (P3H) and lysyl hydroxylases (LH). 4-hydroxyprolines are most abundant modified residues in collagen chains. The appearance of hydroxyproline is approximately 100 4-Hyp-residues in 1000 residue long collagen chain. Proline residues are often hydroxylated in the Y position of the repeat Second most hydroxylated residue is lysine residue varying by collagen type ranging up to ~70 Hyl residues per 1000 residue chain. Both 4-hydroxyproline and hydroxylysine have the crucial stabilizing effect of collagen stability. 4-Hyp residues tend to have affected more on the stability. Hyl residues ads structural strength and determines the chemical properties. 3-Hyp usually appears in the X-position (Hudson & Eyre. 2013). The number of 3-Hyp residues is the smallest in

collagen chain, with only ~10 residues per 1000 residue chain (Marini *et al.* 2007, Myllyharju & Kivirikko. 2004, Myllyla *et al.* 2007b).

2.4 Structural properties of prolyl-4-hydroxylases

The prolyl-4-hydroxylases may be the most versatile members of the 2-OGDD superfamily. The catalysis of prolyl-4-hydroxylases is part of the most prevalent post-translational modification in humans, which yields (2S,4R)-4-hydroxyproline. Hydroxyproline was first identified and isolated from gelatin hydrolysates by Emil Fischer (1902). The effect of adding hydroxyl group to the proline pyrrolidine ring is very moderate and has a minimal change in the molecular mass of the molecule. However, the modification of single residue can have significant effects in a signaling pathway, system or on a protein. The prolyl 4-hydroxylase substrates are summarized in (**Fig.3**).

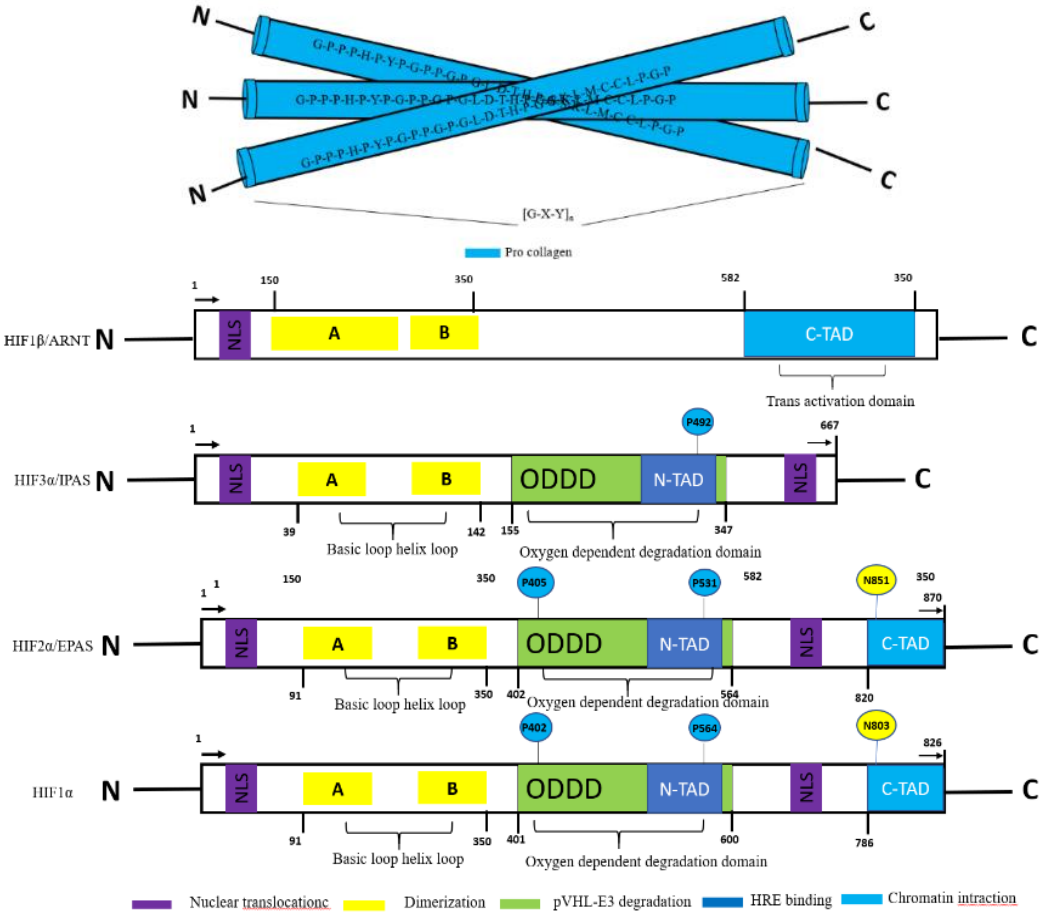


Fig.3 Schematic representations of human prolyl 4-hydroxylase substrates. Procollagen chains with $[G\text{-}X\text{-}Y]_n$ motifs and different isoforms of HIFα and the HIFβ, which binds to all HIFαs.

The most recognizable human prolyl-4-hydroxylases are the HIF prolyl-4-hydroxylases and collagen prolyl-4-hydroxylases. Both enzyme groups share the catalytic mechanism and the DSBH-core fold, but they differ both in size, topology, and substrate specificity. HIF-P4Hs cannot hydroxylate collagen chains, and C-P4H cannot hydroxylate HIF1α like peptides in animals (Jaakkola *et al.* 2001, Kant *et al.* 2013) The collagen triple helical structure forms only when 4Hyp stabilizes the α-helix coils. The single most significant regulation process in HIF-regulated hypoxia signaling is dependent on the hydroxylation of HIF.

2.4.1 HIF prolyl-4-hydroxylases

The reaction mechanism of high affinity towards oxygen enables HIF prolyl-4-hydroxylases to act as an oxygen sensor. There are three known isoforms of HIF-P4Hs (HIF-P4H1, -2, -3,). Often the nomenclature of these enzymes is different in the literature depending on the source. The first evidence HIF-VHL mediated HIF1α degradative pathway was described in *C.elegans* conserved nematode pathway, where interaction between HIFα and pVHL was recognized to be hydroxylation dependent (Epstein *et al.* 2001, L. E. Huang *et al.* 1998). Further studies revealed the hydroxylated residues proline residues in ODDD which is the recognition site for the pVHL-E3 ubiquitin ligase (Ivan *et al.* 2001, Jaakkola *et al.* 2001). Based on the of nematode *C. elegans* Egl-9 gene, which presented egg-laying defect phenotypes, those human homologs were named as egg-laying deficient nine – like protein, EGLN1,-2,-3 (M. S. Taylor. 2001). Another interpretation of the enzyme name refers to the prolyl hydroxylation capability of HIF, hence the name prolyl hydroxylation domain enzymes, PHD -1,-2,-3 (Bruick & McKnight. 2001).

The significance of oxygen has been demonstrated with kinetic experiments, mainly measuring the Km of HIF-P4Hs towards oxygen. The kinetic data of HIF-P4H has shown that the enzyme acts as a sensitive oxygen sensor. The Km values for O_2 has ranged from various *in vitro* experiments from ~0.5-2.5x10^2μM, and most recently even up to 4.5x10^2μM range (Hirsila *et al.* 2003, Koivunen *et al.* 2006, Pektas & Knapp. 2013). The HIF-P4Hs oxygen Km value ranges from 230-250μM and even 100μM depending on the HIFα substrate analog size, but notably higher than the recombinant HIFα ODDD (Koivunen *et al.* 2006). These values are very high compared to the physiological concentration of oxygen in the cells 10-30μM depending on the tissue. However, the oxygen level in the cells may never reach the maximum

value the *in vitro* experiments demonstrates that the capacity of HIF-P4Hs is high. Therefore, the hydroxylation of HIF is considered as the rate-limiting step of the hypoxia response. The sensitivity makes HIF-P4Hs an active reporter of O_2 concentration within the physiological range (Schofield & Ratcliffe. 2004).

The core fold of all HIF-P4Hs resembles the distinguishable double-stranded β-helix fold. The crystal structure of HIF-P4H2 catalytic domain starting from Pro181 was the first HIF pathway enzyme solved (McDonough *et al.* 2006). The structure revealed the distinctive DBSH-core fold with the 2-His-1-Asp facial triad incorporating the Fe^{2+} 6-coordination with bivalent 2-OG. The binding poses of the catalytic residues also supported the reaction mechanism, where the 2-OG carboxylate and ketone group forms coordination bonds with Fe^{2+}. The binding site located in the interface between the β5-β7 strands, and to the residues H313, H374, and D315. The 2-OG orientation forms multiple polar contacts from R383, T329 and surrounding water molecules (McDonough *et al.* 2006, Rosen *et al.* 2010a). The binding pocket flexibility is caused by the β2-β3 loop region (β2-β3 being part of the DSBH), allowing the primary substrate to enter the core. Based on the sequence, HIF-P4Hs main differences are in the N-terminal domain, and there are no available crystal structures. HIF-P4H1 and -3, N-terminal regions have either more disordered structure or lack the extension. HIF-P4H2 N-terminal regions contain an MYND-type zinc finger domain, also appearing in higher animals HIF-P4H.

The findings from the HIFα and pVHL-Elongin C-B complex show that the Hyp residue favors C4-*exo* conformation when in the native form the proline residue is in C4-*endo* conformation. Substrate analog structures with HIF-P4H2 has shown to induce conformational movement in the β2β3-loop, and the water molecule solvation is very different. The reaction rate of HIF-P4H2 is moderately slow compared to the studies done in viral P4Hs and TauD (Flashman *et al.* 2010, Longbotham *et al.* 2015). These suggest that the protein may undergo some extensive reconfiguration before the substrate oxidation and requires stabilization before O_2 can bind. The buried location of the catalytic site inside the core and the stabilizing binding has effects to the kinetics. The activity of HIF-PHD2 is not reduced even without the addition of endogenous iron. The activity supports the proposal that the iron-facial triad coordination is very stable, and the reaction rapidly proceeds when oxygen binds. The slow reaction is contrasted with the high

O_2 Km values, suggesting that, the enzyme is available for substrate oxidation rabidly which rationalizes the HIF-P4H role as an immediate hypoxia sensor though HIFα hydroxylation.

The structural biological knowledge of HIF-P4H to HIF1α binding has prompted the forward the structural based drug discovery of, in order to find potential P4H-inhibitors. Today there are 67 protein data bank entries with ODDD substrate analogs or with HIF-P4H inhibitors.

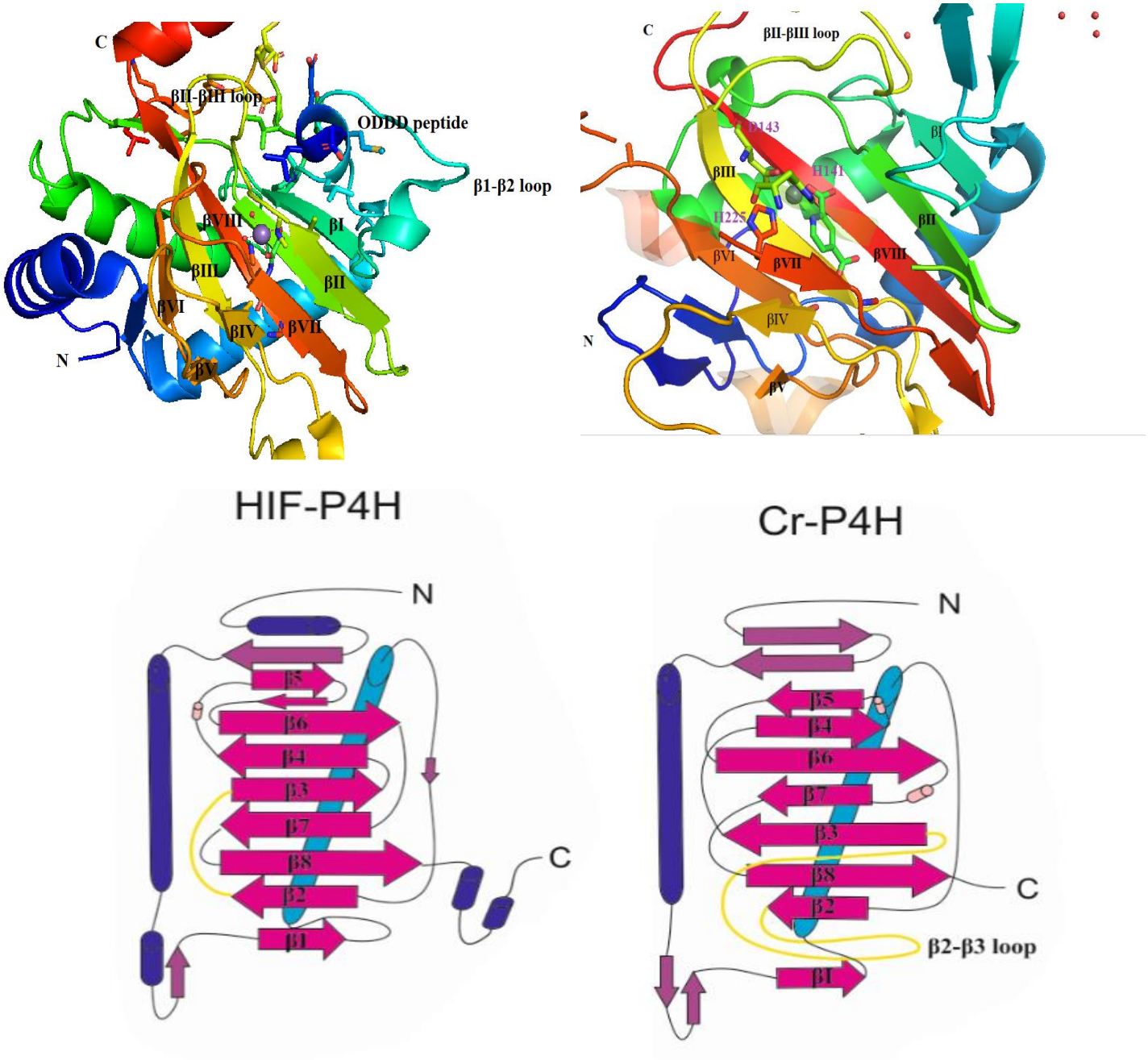


Fig.4 Crystal structure of HIF-P4H2 bound to Mn^{2+}-cation with 2-OG analog N-[(4-hydroxy-8-iodoisoquinolin-3-yr)carbonyl]-oxalyglycine and HIF-ODDD peptide DLEMLAPYIPMDDDFQL (**A**). The sheet ran in an antiparallel pattern linked by loop regions

and supported with two N-terminal helices. The β2-β3 loop-helix which undergoes a conformational change during substrate binding. The analogs induce a conformational change in the β2-β3 loop from open form to closed form towards the core fold. Crystal structure of Cr-P4H monomer highlights the differences of P4H enzymes (**B**). The main difference is the extension of the βII-βIII loop, which is longer in C-P4H. The second difference is the shorter and unstructured C-terminal loop. In a way the C-terminal loop and βII-βIII loops have inverse roles in these P4Hs, which may have a role in substrate specificity *in vivo*. The topology maps simplifies the differences (**C**). Coordinates were taken from PDB entries 3HQR and 2JIG (Chowdhury *et al.* 2009, McDonough *et al.* 2006, Rosen *et al.* 2010a)

2.5 Collagen prolyl-4-hydroxylases

The maturation of collagen chains is dependent on the function of collagen prolyl-4-hydroxylases (C-P4Hs). In humans, there are three isoforms of the enzyme forming ~240kDa α2β2-heterotetramers. Conserved tetrameric homologs with similar function, is seen in other vertebrates, nematodes, insects, and plants such as *C.elegans, D melanogaster* (Hieta & Myllyharju. 2002). C-P4Hs catalyze the proline hydroxylation of collagen polypeptide chain $[G\text{-}X\text{-}Y]_n$-repeats where the 4-hydroxyproline residue locates in the Y-position. Together with the other important post-translational modifications lysyl-5-hydroxylation and proline-3-hydroxylation, collagen chains form the folded functional triple helical structure. The 4R,2S-isomer of Hyp, has a critical role in stabilizing the triple helical structure. The amount of Hyp in the Y-position varies, and not always satisfy the motif sequence. Eventually, all proline residues are hydroxylated in the mature collagen (Myllyharju. 2003).

Humans have three isoforms of each collagen modifying prolyl-hydroxylase. Human C-P4H isoforms differ mainly by their α-subunit structures. Like in HIF-P4Hs, C-P4H α-subunit sequence identity is divided so, that isoforms I (α-I) and two (a-II) share more conserved residues respect to isoform II (α-III). The α-I and α-II share 65% sequence identity, whereas α-III share only 35-37% identity with the later ones (Annunen *et al.* 1997, Kukkola *et al.* 2004). The conservation is much higher in the C-terminus among the isoforms — the ratios of α-1/αII and α1α2/α3 shares 80% and 56-57% identity respectively (Kivirikko & Pihlajaniemi. 1998, Myllyharju. 2003, Myllyharju. 2008).

The vertebrate C-P4Hs consist of two identical catalytic α-subunits and β-subunits. The β-subunit is identical protein disulfide isomerase (PDI) which belongs to the thioredoxin superfamily member, which is an ER molecular chaperone promoting protein folding (Pihlajaniemi *et al.* 1987, Tasanen *et al.* 1988, Vuori *et al.* 1992). The β-subunit is identical to the thioredoxin superfamily member, protein disulfide isomerase (PDI), which is an ER molecular chaperone promoting protein folding. In the ER, PDI catalyzes the formation and isomerization of disulfide bonds utilizing glutathione as a reducing species. PDI is a heterodimer consisting namely **a, a', b** and **b'** subdomains taking distinguishable thioredoxin-like fold (Wallis *et al.* 2009). The catalytic site of PDI family proteins is well understood, but the binding properties of the substrate in controversial and many of the PDI family members are well characterized. The **a** and **a'** subdomains catalyzes the disulfide bond formation, and **b** and **b'** are responsible for substrate binding. It has also shown that **a'** may take part in the substrate binding, but the interaction is very weak compared to **b** subunits (Ellgaard & Ruddock. 2005).

The catalytic site of PDI consist of CXXC-motif. The mobility shift of nearby arginine residue induces pKa shifts of the two catalytic cysteine residues varying the residues between thiol and thiolate states. The change in the pKa of these residues induces nucleophilic attack to the glutathione thiol groups cycling reduced and oxidized states of thiol groups. The donated proton from glutathione eventually causes substrate cysteines to oxidize to a disulfide bond. The same mechanism is present in the isomerase activity allowing rearrangement of the disulfide bonds. In C-P4H the role of PDI is controversial. It is assumed to retain the highly insoluble catalytic domain in soluble form and to localize the enzyme in the ER with the [KDEL]-ER retention sequence (Kemmink *et al.* 1996). There is no evidence that PDI catalyzes disulfide bond formation in C-P4H, suggesting that PDI not interact actively with the α-subunit (Koivunen *et al.* 2005).

The PSB-dimerization domain assembly is referred as the double-double domain, connecting to the CAT-domain with a 70-residue long loop linker which is the first characterized structure in human C-P4H ((Hieta & Myllyharju. 2002, Hieta *et al.* 2003). The PSB (~100 residues) domain consist of tetratricopeptide repeats, which appear in protein which form protein-protein, or protein-peptide interactions. The binding groove of the domain is abundant in aromatic residues (Tyr, Phe in PSB), providing a binding surface for the elongated collagen peptide

substrates to bind. The crystal structures of human PSB-I shows that the peptide substrate analog $[PPG]_9$ and $[PPG]_{10}$ binds with consecutive three tyrosine residues in each helix (total six Tyr residues), and a phenylalanine residue at the end. Site-directed mutagenesis pointed out that essential residues for binding are Tyr158, -163, -199 and 233 (Pekkala *et al.* 2004). The structures of human double-double domain are known which brought important insight of the heterotetramer assembly (Anantharajan *et al.* 2013). The dimerization domain consists of four helices α1-α4 extends into a helical bundle, where the first helix intertwines as a coiled-coil structure. The other two helices, α5-a6 closes the motif and provides stability of the PSB-domain contact. The two domains link to each other with ten residues long loop followed by a short linker (Murthy *et al.* 2018a).

The catalytic α-subunit of C-P4H is composed of three subdomains, the catalytic CAT-domain, peptide-substrate binding domain (PSB) and the N-terminal dimerization domain linking the homodimers together (Hirsila *et al.* 2003). Since the structural information about C-P4H has beveled just over a decade, a large extent of work has done utilizing in vivo mouse models and kinetic experiments. The hydroxylation of polypeptide chains is rather well understood relative to the structural information available of the catalytic site (Hirsila et al. 2003). The position of the residues in the [X-P-G]-polypeptide has dependence in the hydroxylation rate and efficiency. C-P4Hs are not able to hydroxylate [G-X-P], or [P-G-X]-peptides in vitro. Likewise, in HIF-P4H CTAD peptides, the length of the collagen-like peptide affects the hydroxylation. The C-P4H1 Km towards the peptide tends to decrease when the peptide is extended. Similar results have been observed in other two C-P4H isoforms. The proposal of the full-length C-P4H is presented in (**Fig.5)**

In recent understanding, the CAT-domain is thought to have the DSBH-core fold, according to sequence analysis and available crystal structures from lover organisms algal C-P4H *Chlamydomonas reinhardtii* (Cr-P4H), which has 27% sequence identity with the human enzyme (Lieb *et al.* 2002). The structure possesses the distinct DSBH-fold with eight antiparallel β-sheets, with a longer βII-βIII-loop than in HIF-P4H2 (see fig.4) (Koski *et al.* 2009a). The peptide substrate analog is bound in the enzyme within the groove formed by β3-β4 and βII-βIII loop where the catalytic residues are orientated in the middle of this groove. The binding motif of the peptide substrate resembles the binding of HIF1α, with a 2-His-1-carboxy facial triad in complex with Fe^{2+}. Other conserved residues in the C-P4H group is the

lysine and arginine residues binding to 2-OG carboxylate C5 atom. The significant difference in the hydroxylation is the uncoupled relay reduction of the iron complex, by L-ascorbic acid. The necessity of ascorbate is quite not understood, and the enzyme is active in the uncoupled 2-OG decarboxylation, without the native peptide substrate hydroxylation. The Uncoupled reaction is not inhibited even with saturated substrate concentrations. The apparent reason for the uncoupled reaction is not clear, but the *in vivo* experiments and ascorbic acid deficiency-related diseases has shown that it is a vital part of collagen synthesis.

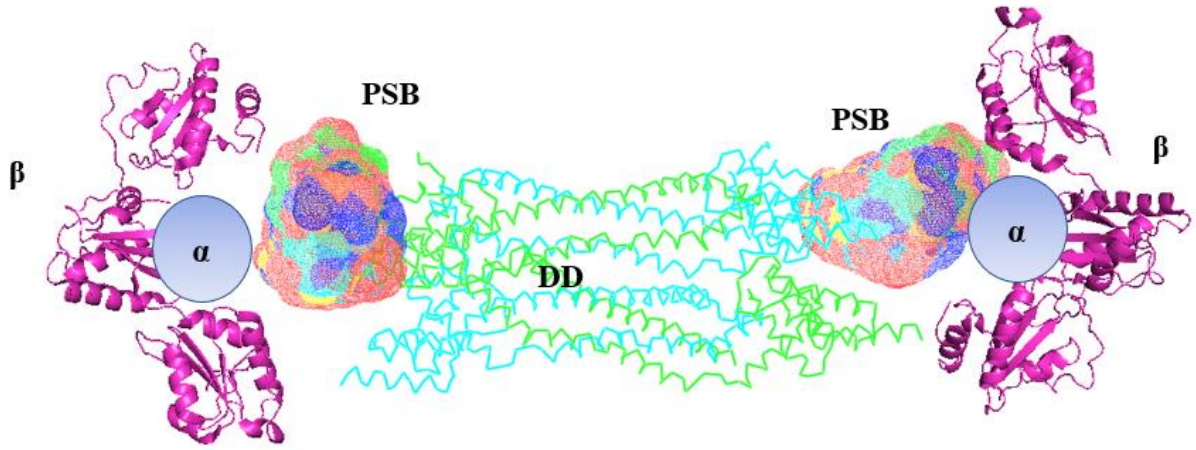


Fig.5 Current proposal of the C-P4H enzyme. This model is proposed based on the SAXS shape models and available crystal structures. Currently, the crystal structures of the PSB domains and dimerization domain is known and in this model PDB entries 2YQ8 and 6EVL (Anantharajan *et al.* 2013, Koski *et al.* 2017, Murthy *et al.* 2018b). Also, the β-subunit of the CAT domain is known through PDI structures. In this model PDB entry 3UEM (Wang *et al.* 2012). However, the complete structure or the detailed positions of this domain is not known. The model is based on the recent schematic published (Koski *et al.* 2017).

2.6 Transmembrane prolyl-4-hydroxylase

The most recent member identified prolyl-4-hydroxylase is the ER membrane transmembrane prolyl-4-hydroxylase (P4HTM). The gene is found mostly in vertebrates, *Danio rerio* sharing 51% gene sequence identity to human. Some exceptions appear in some anthozoans, but it is found in flies or nematodes. P4HTM is a ~50 kDa enzyme localized in the ER lumen with [DARVEL]-signal sequence, and it is bound to the membrane with 24 residues short transmembrane region (Koivunen *et al.* 2007). The main fold should refer to as the 2-OG

dioxygenase domain. There is clear evidence that the key features of DSBH-fold elements are present, but the unavailability of any data-based model restricts this interpretation. However, the enzyme activity is dependent on the 2-OG and iron co-factors, stating that it is indeed a 2-OGDD member. It is expressed in significantly higher levels in the brain, adrenal cortex and retina and a lower amount for example in skeletal muscle and digestive track. However, the localization of the enzyme indicates that may not at least immediate mediator of HIFα (Koivunen *et al.* 2007).

P4HTM seems to now have specified substrate binding domain, but that the substrate is bound similarly as to HIF-P4Hs. Based on the sequence analysis the main enzyme fold most likely adopts the DSBH-fold with variable converged loop regions. These features are conserved in P4HTM especially regarding Cr-P4H. Also, P4HTM has three potential N-linked glycosylation sites in asparagine residues N382, N368, and N348. The work done on two different constructs *in vitro*, full-length and transmembrane truncated D88-L502 form has found to be homodimers in mammalian and insect cell purified fractions (Koivunen *et al.* 2007).

The sequence identity corresponds 14-15% to C-P4H α-I and α-III, in and in turn 12-13% of HIF-P4Hs 1-3. The catalytic 2-OG and iron binding residues are conserved, and the polar contact to 2-OG is lysine like in C-P4Hs. The disulfide bond between the βVII-βIV loop and βVI-sheet from Cr-P4H, which is also present in P4H-TM (Koski *et al.* 2009).

2.6.1 Clinical relevance of transmembrane prolyl 4-hydroxylase

The distinctive role of P4HTM is unambiguous. For one reason, it arises from the localization of P4HTM respect to HIF-P4Hs. The studies done one HIF-P4H indicates that HIFα is hydroxylated in the cytosol and transported to the nucleus. Therefore, the localization of P4HTM does not fit to these findings. Cellar experiments showed that under hypoxia P4HTM is upregulated. Mammalian and insect isolation extracts have shown that P4HTM seems to decrease amounts of HIFα and suppression of the P4HTM gene with siRNA results increasing levels of HIF1α. P4HTM can hydroxylate HIF ODDD peptide analogs *in vitro,* but do not hydroxylate collagen peptides (Myllyharju & Karppinen. 2013). However, mutated HIF ODDDs peptides where the target prolines are mutated to alanines (J. Huang *et al.* 2002, Koivunen *et al.* 2007).

Studies done in transgenic animals *in vivo* have shown that knockdown of P4HTM gene resembles severe defects in basement membrane construction and kidney function in zebrafishes. The morphological changes appeared in the head size in eyes which especially contains basement membranes. The mRNA levels of HIF-target genes supported the phenotypes, where some ECM components, such as *Lox* was decreased and type XVIII collagen was increased. These findings suggest that P4HTM may have other substrates that are not directly associated with the HIF-pathway (Hyvärinen *et al.* 2010). The link between HIF-pathway and P4HTM was the observation, that P4HTM also contributes the EPO regulation *in vivo* in mice. The established mouse models for HIF-P4H knockouts have shown that the HIFα stabilization occurs inducing EPO production, HIF-P4H2 being the most prominent regulator. The HIF-P4H target inhibition with administered drugs also shows similar effects. Similar results were observed with P4HTM, with the exception that EPO levels appear to be increasing in kidney, but not in the liver. This suggests that P4HTM has at least a contributing role to EPO production and also another HIF-target gene expression (Laitala *et al.* 2012).

More recently the defects found in zebrafish eyes and retinal base membranes was supported with similar results found in P4HTM knockout mice. Similar genes (*Lox* and *Pai-1*) were upregulated in brains and kidneys respect to zebrafishes, but not significantly in the eye. HIF1α or HIF2a was not stabilized in the brain, eye or kidney in P4HTM. However, significant differences were found in the retinal pigment epithelium. The outer nuclear layers where the primary photoreceptors reside was found to be significantly thinner or even damaged. Since the role of retinal pigment epithelium is to recycle the photoreceptors, the retinal function is compromised due to the knockout (Laitala *et al.* 2012).

Additionally, the same model revealed also increased EPO levels despite the unobserved HIFα stabilization. Since recent discoveries P4HTM can hydroxylate HIFα ODDD, thus including to the HIF-mediated hypoxia response (Laitala *et al.* 2012). It seems that the P4HTM has a more supporting role regarding the HIFα stabilization associated gene regulation, especially in those tissues where sufficient HIF-P4H expression is reached. The contribution to EPO regulation also supports this statement. However, P4HTM has an essential role in specific tissues, as described especially in developmental level. From a clinical aspect, P4HTM gene mutations, which have been identified as a cause of severe intellectual disabilities. There has been reported a Norther-Finland family which syndrome is causing a delay in the intellectual development, in

which one of the suspected genes are P4HTM along with transketolase (TKT) and ubiquitin specific peptidase 4 (USP4) (Kaasinen *et al.* 2014, Leinonen *et al.* 2016).

2.7 X-ray scattering provides information about protein structure

2.7.1 X-ray crystallography

The concept of X-ray scattering involves the interaction between the electromagnetic wave with matter resulting in diffraction. The diffraction can be displayed as a pattern.

X-ray scattering is a technique used in structural biology that has been available for over 100 years. When X-ray beam hits macromolecule such as protein or nucleic acid, the wave diffracts, resulting in a pattern. In order to cause diffraction from the atom, the wavelength of light needs to be the size of an atom and to maximize the sampling, a large number of molecules is needed. For this purpose, the sample needs to be crystallized. X-rays are necessary for structure determination because the interparticle distances in the molecules are very short. The diffraction is the result of photons from the X-ray source colliding with the electrons clouds of the molecule. The electromagnetic field of the photon causes oscillation in the electrons of all the atoms. The oscillation creates a dipole moment and coherent radiation which is elastic scattering. The energy and the wavelength of the scattered wave do not change, when the scattering is elastic. The electromagnetic wave is composed of two vectors, the electric field *E* and the magnetic field vector *B*, which is perpendicular to the *E* field (Ryland. 1958). The cross product of these vectors results is the wave vector S_0. Depending on the wavelength of the photon, the magnitude of the wave vector is the inverse of the wavelength. The basic concept of X-ray diffraction is brilliantly interpreted in the Bragg equation [1], where the scattering vector is switched to point up. Now the scattering of incident beam S_0 can be seen to reflect from the crystal plane. This leads to the relation between the scattering angle Θ and distance of the two scattering centers. This then tells the multiple integers for maximum wavelength

$$n\lambda = 2dsin\,\Theta \text{ [1]}$$

All the atoms in the specimen contribute to the scattering amplitudes of the scattered waves, which can be either in phase or out of phase, or constructive or deconstructive. Because, there are no focusing lenses to direct the beam, mathematics is used to determine the position of atoms in the molecule. The atoms are packed inside a unit cell. It is the smallest single unit inside the crystal. Coordinates of each atom can be described with a tree spatial coordinates

x,y,z using the vertex or lattice point as a reference also called the bravais lattice (Huggins. 1945, McPherson & Gavira. 2013). The reflections that are scattered in a space, which is called the reciprocal space, usually denoted with h, k, and l for each atomic coordinate assigned as integers. Thus, the crystallographic data collection consists of two parameters; the reciprocal space coordinates hkl and the intensity of the reflection I_{hkl}. The mathematics are required to treat the sampled reciprocal space. Fourier transform is a continuous integration method to combine sine and cosine terms of the wave to construct a complex wave. The sum of each scattered scattering factor is described with the structure factor equation, which can be simplified in eq.1 [2]

$$F_{hkl} = f_A + f_B + \cdots f_{A'} [2]$$

The structure factor implies that each reflection from the unit cell, of many waves resulting in diffraction by individual atom. Because the scattering is scattered from the electron orbitals, not the nucleus, the structure factor equations take for eq.3, where the coordinates are expressed. [3]

$$F_{hkl} = f(\rho 1)_A + f(\rho 2)_B + \cdots f(\rho m) + f(\rho n) \text{ [3]}$$

In macromolecular crystallography, the amplitude and the frequency of the Fourier term are measurable. However, the phase is not. One could look the phases on an index, in case of the more simple inorganic compound. In order to get the phases of a protein molecule, ρ(x,y,z), in reciprocal space must be calculated. The mathematics of the Fourier series is rather complicated and not derived here. In general, the Fourier transform is integration method where any function *fx* can be presented as *F*(*x*) where F(*h*) is the Fourier transform of f(x), and the variable h is in reciprocal units. The structure factor equals the vector sum of all scattering vectors from all the atoms in the unit cell. The Fourier series of structure factors is the integral over the unit cell volume. Since the reciprocal space is three dimensional, the integral should be taken from an equal number of volume, or equivalently the total volume *V* [3]. There are two notations for structure factor, F_{hkl} and S(q), which is often referred to as the Debye scattering equation.

$$F_{hkl} = \int_v \rho(x,y,z) e^{-2\pi i(hx,ky,lz)}\, dV \text{ [4]}$$

Since the Fourier transform is reversible and the electron the electron density is a discrete variable. The electron density can be expressed as a triple summation. The requirement of expressing continuous integral is that the sum is an infinite sum of infinitesimal, which can be then used to compute the phases, where the Fhkl is a vector amplitude with a phase α eq.4 [4] (Rupp. 2010).

$$\rho(x,y,z) = \frac{1}{V}\sum h \sum k \sum l \; F_{hkl} e^{-2\pi i(hx,ky,lz)} [4]$$

The simplified description of the phase problem tells that taking the Fourier transform of ρ(x,y,z), the phase α can be calculated and thus used to construct a model of the protein. It is also possible to represent the F_{hkl} in scalar form when it can be used as a complex plane. There are different methods to approach this problem. The most often and easiest method is to use the molecular placement method, where the phases from known homolog structures can be used to approximate the phases of the target protein. The density maps of the phasing modeling can be then iterated taking the fraction of the two structure factors. This is called refinement. However, if there are no available homologs, the has to calculated with another dataset using heavier atoms., there are multiple computational approaches to obtain the phases from the heavy atom dataset. Most well-known methods include multiwavelength anomalous dispersion (MAD), multiple isomorphous replacements (MIR), and single wavelength dispersion (SAD) (Blaho. 2008, Hendrickson & Ogata. 1997, G. L. Taylor. 2010).

2.7.2 Small-angle X-ray scattering (SAXS)

Small angle x-ray scattering as a structural determination method is used to characterize particles in the solution. X-rays scattering pattern from solutions is isotropic, and the measured intensity depends only from the momentum transfer vector, which is denoted depending on the interpretation as *s* or *q* eq.5 [5].

$$q = \frac{4\pi \, sin\theta}{\lambda} \; [5]$$

SAXS measurement is always a comparison measurement. Because the system is in solution and not in a crystal lattice, the number of scattering centers are constantly moving. Therefore, the scattering pattern is an average of the sample matrix, and the scattering pattern is isotropic. The intensity of the sample is defined, by taking the ratio between the scattered energy towards the area 2Θ, which is denoted as I_0. Given that the system is isotropic and assuming that the solution is monodisperse the intensity of the solvent has to be subtracted from the intensity of the sample. The main difference to the crystal lattice in solution is that the structure factor F does not have dimension h,k,l in the reciprocal space. Because the scattering is sampled from a cross-section of scattering intensity, F is described as the from a factor of the particle *F(q)*. Equally important is to note that now the solution contributes also the scattering, which will

have own structure factor of the solution *S(q)*. Therefore the parameters measured in SAXS experiments can be expressed as Eq.[6](D. Svergun. 2013, D. I. Svergun & Koch. 2003)

$$I_{(q)} = F_{(q)}\, S_q \text{ [6]}$$

Since the particles in the solution are at certain time point apart with a distance *r*, the phase can also be factored, which will take the form to $be^{(Iqc)}$=*sin(qf)/qr*. This is the reciprocal space form factor of the particle distance distribution. It is favorable to integrate over the particle to treat the scattering length density function ρ(*r)* as a continuous integral which is the inverse Fourier transform [7]. This is an integral that is limited to the maximal dimension of the particle. It should be noted that the scattering vectors from the crystal unit cell are defined by the reciprocal lattice $F_{hkl,}$ and the amplitudes are summed.

$$I_q = \int p_{(r)} \frac{Sin\,(qr)}{qr}\, dr \text{ [7]}$$

However, in SAXS measurement, it is more meaningful to refer as sums of scattering intensities, because there is no diffraction pattern. The information that the distance distribution function formulates is the shape and dimensions of the particle. The main disadvantage of SAXS is the absence of atomic-level resolution, but it is a very powerful technique to determine conformational properties, domain dynamics, and even behavior of mixtures. The most important quantified parameters from SAXS measurement are the maximum particle distance (Dmax), the radius of gyration (Rg) and molecular weight. Rg and Dmax quantitative values calculated from the real space distance distribution curve and from the Guinier analysis. The intensity of the scattering centers increases as a function of decreasing angle. The Guinier approximation consists a linear part in the scattering curve which allows approximating the zero angle intensity (I_0). The notation was formulated by Andre Guinier. The quality of the data and the matrix disturbances caused by aggregation or repulsion from the buffering solution will make the analysis impossible.

3. AIMS OF THE STUDY

This study aims to determine the structure of P4HTM constructing molecular model either using MX of SAXS. European beamline sites are used to collect high-quality data. In order to support this goal, the truncated TmΔ88-502 variant is characterized using spectroscopic techniques (CD, SEC-MALS, DLS). In order to ensure the biologic relevancy, the activity of the enzyme is determined with uncoupled P4H 2-OG decarboxylation assay (Kivirikko & Prockop. 1972). Baculoviruses encoding shorter P4HTM constructs are generated suitable for *Spodoptera frugiperda* strain 9 protein expression. The novel constructs are generated in order to maximize the probability to obtain a sufficient amount of soluble protein for structural experiments. The thesis should be able to answer the basic functional feature of this protein and point out the important features, such as stability, secondary structure, important post-translational modifications, and co-factors.

4. MATERIALS AND METHODS

4.1 Bioinformatics

Two modeling platforms were used to predict the assumed full-length P4HTM structure and the secondary structure distribution. The template search and model building was done using two inputs, the full-length P4HTM 1-502, and a TmΔ88-502 truncated variant. The programs used for the modeling was Phyre2 (Kelley et al. 2015) and SwissModel (Benkert et al. 2010, Bertoni et al. 2017, Bienert et al. 2017, Guex, Peitsch et al. 2009, Waterhouse et al. 2018). The full length and the TmΔ88-502 truncated variant were run ten times. The first five input cycles had a more straightforward annealing mode with fewer iterations and the remaining five with more intense and slower annealing mode. Both programs yielded similar results in each cycle which were evaluated with Pymol and selected for merging the models in the server interface (Schrödinger. 2018). The coverage and the confidence of the model are given as a percentage (Guex, Peitsch et al. 2009).

Results obtained from the template search was used to construct multiple sequence alignment table for the full-length P4HTM. The multiple sequence alignment inputs were chosen so that,

the alignment contains P4Hs from all different species, with at least sequence identity >20%. The selected alignment inputs were human HIF-P4H, human C-P4H-αI, Cr-P4H, Ba-P4H, and PBCV-1. Additionally, a similar alignment was done for the N-terminal part of P4HTM, containing N-terminus, a transmembrane helix and the EF-hand motif. The primary alignments were made using the Clustal-W algorithm with Clustal-X program (Larkin et al. 2007, Thompson et al. 1994). The sequence editing and secondary structure depiction were done using STRAP (Gille & Frömmel. 2001), GENEDOC, and ESPript (Robert & Gouet. 2014). The transmembrane topology prediction was made using HMMTOP, and DAS transmembrane prediction algorithms (Cserzo *et al.* 1997, Tusnady & Simon. 1998) sequence was put through, which resulted in a single helix through the membrane

4.2 Protein construct design and expression vectors

Four truncated constructs of the P4HTM isoform one was designed. All of the constructs lacked the transmembrane domain. All of them were designed to be shorter than the available $P4HTM_{D88\text{-}L502}$ construct described by Koivunen and colleagues (2007). Structure of truncated P4HTM was modeled using the web-based protein modeling server Phyre2 (Kelley *et al.* 2015). In shortening order, the additional constructs were E115-, D130-, G149- and Q187-L502, based on the protein model secondary elements order. The reading frame starts with a PDI signal sequence, (MLRRALLCLAVAALVRA), followed by N-terminal 6xHis-tag in all constructs. The existing pVL1932 (Invitrogen) baculoviral vector was used as a template to generate the new PCR-products. Additionally, for bacmid based baculovirus generation, pFastBac™-Dual (Invitrogen) were used.

4.3 Cloning and baculovirus generation

The coding sequences of P4H-TM variants was generated for two baculovirus compatible transferring systems. The pVL1392 vector contains homologous regions for the AcMNPV baculovirus genome when co-transfected with the pVL1392 results homologous recombination in the genome in the transfection phase. An additional approach is to use the Bac-to-Bac system where the gene of interest can be cloned, into the baculovirus genome is already included in bacmid for this purpose pFastBac™-Dual vector was used.

The pVL1392 vector inserts were generated using a site-directed mutagenesis approach. The mutagenesis was done using the available pVL1392-P4HTM$_{D88\text{-}L502}$ plasmid as a template. The sequences were amplified in reverse 3'-5' direction where coding sequence ends respectively before D88 including the ER retention signal (5'GATGCGCGCGTGGAGCT 3'). The pFastBac™-Dual vector inserts were generated using restriction-digestion compatibility. The oligos used for PCR and sequencing purposes are summarized in (**Table.I**)

Table.I The oligos used for construct generation[a] and sequencing[a].

Oligo	Sequence
pVL1392$_{E115\text{-}L502}$[a]	5'GAGGGCATCAAGGTGGG 3'
pVL1392$_{D130\text{-}L502}$[a]	5'GACAGGGATCACTTCATCCGA 3'
pVL1392$_{G149\text{-}L502}$[a]	5' CAGGTCAGCCAGCTGGA 3'
pVL1392$_{Q187\text{-}L502}$[a]	5' CAGGTCAGCCAGCTGGA 3'
pVL1392$_{Reverse}$[a]	5'ATGATGATGATGATGATGGGC 3'
pFastBac$_{SalI\text{-}forward}$[a]	5' TTAAGTCGACCATGCTGCGCCGCG 3'
pFastBac$_{XbaI\text{-}reverse}$	5' TTTATCTAGATCAGAGCTCCACG 3'
M13/pUC $_{Forward}$[b]	5' CCCAGTCACGACGTTGTAAAACG 3'
M13/pUC $_{Reverse}$[b]	5' AGCGGATAACAATTTCACACAGG 3'
Polyhedrin $_{Forward}$	5' AAATGATAACCATCTCGC 3'

The mutated pVL1392 plasmid was generated using Biorad T100 thermocycler. The PCR reaction mixture final concentrations was 1x Phusion HF buffer (Thermo-Fischer), 0.2 mM dNTP mix, 0.5 mM forward primer, 0.5 mM reverse primer, 0.02 U/µl Phusion polymerase (Thermo-Fischer) and two ng of the pVL1392$_{D88\text{-}L502}$-template. The PCR-products was running on 0.8% agarose gel in 40 mM Tris, 20 mM acetic acid, one mM EDTA-buffer for 80V/1h. The gel was visualized using Biorad ChemiDoc™ XRS+-gel imager. The PCR product size was expected to be approximately 11 kb with blunt cohesive ends. The gel pieces were cut on UV-board and dissolved into sterile water. The Product was purified using Illustra GFX PCR and gel band purification kit (GE healthcare). The blunt ends were phosphorylated using T4 polynucleotide kinase in 70 mM Tris-HCl (pH 7.6), 10 mM MgCl2, 5 mM DTT-reaction buffer (New England Biolabs). The total reaction volume of 20µl was incubated for 1 hour at +37℃. The plasmid was ligated using T4 DNA ligase (New England Biolabs), in the previous buffering condition for 40 min at +37℃.

The ligation reaction transformed into *E.coli* DH5α competent cells containing ~50 ng of ligated plasmid. The cell suspension was incubated 30 min at room temperature and heat shock

45 sec. +42 ℃ and incubated 2 min. On ice. The transformed cells were incubated +37/180 RPM for 1 hour in 2% tryptone, 0.5% yeast extract, 10mM NaCl, 2.5mM KCl, 10mM $MgCl_2$, 10 mM $MgSO_2$, 2mM D-glucose-medium (SOC). Cell suspensions was inoculated into 4% Tryptone, 2% yeast extract, 172 mM NaCl (LB), 1.5 % agar-plates, containing 100 µg/ml ampicillin. The plates were incubated +37 o/n. The colonies were analyzed using colony-PCR. The final concentration of the PCR reaction in 20 µl reaction was 0,3 mM Polyhedrin-forward primer, 0.3 mM pVL1392-reverse primer, 0,4 U/µl *Taq* polymerase in 20mM Tris-HCl (pH 8.4), 50mM KCl, 1 mM $MgCl_2$-buffer (Thermo Fischer).

Positive colonies were inoculated into 2ml plasmid preparation LB-medium cultures, containing 100µg/ml ampicillin. The cultures were incubated +37℃ o/n, 225 RPM. Plasmids were isolated from the minipreps using the Nucleospin® plasmid purification kit (Macharey-Nigel). The purified plasmids were verified using a sequencing service at Biocenter Oulu sequencing facilities. The plasmids were transfected using the baculovirus transfection system provided by the Oxford expression technologies. For the transfection, ¼ dilution of baculofectinII transfection reagent and 20ng of flashBAC ultra- viral genome was mixed with the 100 ng of constructed plasmid. The transfection volume was added up with serum- and antibiotic free TNM-FH medium (Sigma), and incubated 20 min at room temperature. The transfection was performed into a 24-well plate (Corning) *Sf9* culture with the confluence of 50 % in serum- and antibiotic free medium and incubated o/n at +27℃. After initial incubation, the transfection was continued for five days in normal TNF-FH-medium, containing 10% fetal bovine serum, 10% symperonic® F68, 20 mM L-glutamine, 100 µg/ml penicillin-streptomycin. The transfection culture was collected, and the virus containing supernatant was inoculated into 100/20mm cellstar® Petri dishes (Greiner), containing *Sf9* cells with 50% confluency.

The pFastBac-Dual inserts were generated the same way as described above, with different PCR conditions. The purified inserts containing 5'-3' overhanging ends was made vector compatible by dephosphorylation and restriction digestion, which was done simultaneously by incubating the insert and the vector at +37℃ for one hour. The reaction consisted of 1 U/µl of calf intestinal alkaline phosphatase (New England Biolabs) in 50 mM Potassium Acetate, 20 mM Tris-acetate, 10 mM Magnesium Acetate, 100 µg/ml BSA-buffer (CutSmart NEB), 1 U/ml of XbaI and SalI-HF (NEB) and 1µg of the insert and the vector. The treated vector and inserts were purified and ligated using T4 ligase as described above. The ligated vector was transformed into *E.coli* DH5α competent cells, inoculated into LB-ampicillin plates. Colonies were analyzed using colony-PCR using Dream *Taq* polymerase. The colony-PCR mixture

contained, KCl, $(NH_2)_2SO_4$ and $MgCl_2$-dream taq buffer (Thermo Fischer), 0.2 mM dNTP, 0.5mM Polyhedrin-primer, 0.5mM pFastbac-primer, 0,4 U/µl dream tag polymerase (Thermo Fischer) and colony suspension. Positive colonies were inoculated, into LB-ampicillin miniprep cultures, and plasmids were isolated, purified and sent for sequencing verification.

The bacmids were generated by transforming the pFastBac™-Dual plasmids into EmbacY *E.coli* competent cells (Bieniossek *et al.* 2008). The transformation was done similarly as DH5α cells with 1ng of plasmid and incubated 4 hours in SOC medium at +37 o/n, 225 RPM. The EmbacY cells were inoculated into LB-agar plates containing, 50µg/ml Kanamycin, 7µg/ml Gentamycin, 10µg/ml Tetracycline, 40µg/ml IPTG and 25µg/ml X-gal. The plates were incubated for two days at +37, to complete blue/white colony screening. Positive colonies were analyzed with colony-PCR. The positive colonies were inoculated into LB minipreps containing 50µg/ml Kanamycin, 7µg/ml Gentamycin and 10µg/ml Tetracycline. The cells were re-suspended in 15mM Tris, 10mM EDTA, 100µg/ml RNase A-buffer, followed by 1% SDS, 0.2M NaOH lysis. The cell debris was centrifuged in several steps (10min, 17000g, RT). The bacmid DNA was precipitated with 100% isopropanol and washed with 70% ethanol. The purified bacmids were analyzed with colony-PCR using dreataq polymerase and M13-promoter primers. The positive bacmids were transfected into *Sf9* cells with the similar protocol as above, without the flasBAC ultra virus mixture. In addition to the pVL1392 system, the Bac-to-Bac systems transfected cells were visualized using a UV-light microscope (**Fig.6**). The resultant virus transfectants (P1) was then further amplified (P2) before protein expression.

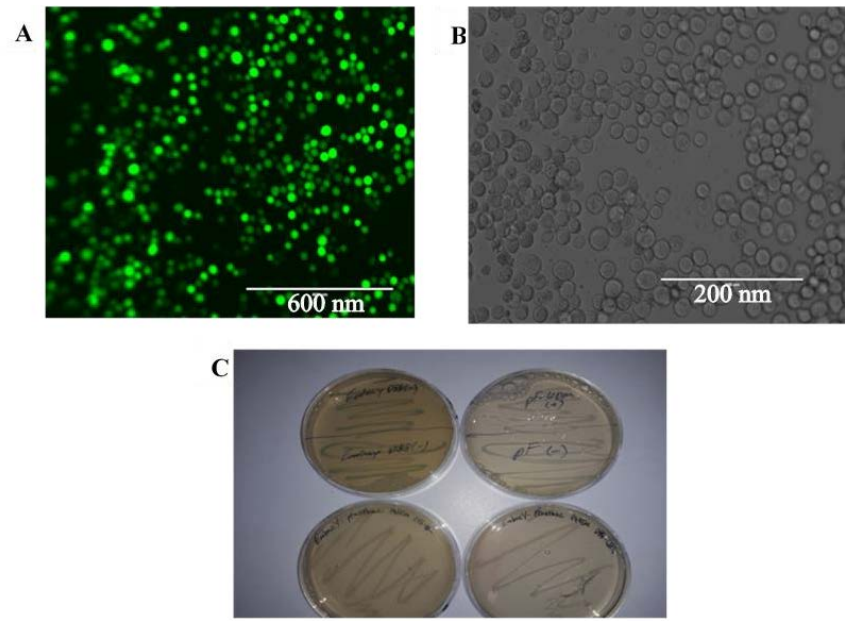


Fig.6 Cloning using the Bac-to-Bac system. The Initial insert is generated using DH5α cells and then transfected into EmbacY baculovirus genome containing *E.coli* strain. The EmbacY features promotor sites for LacZ operon, which enables the blue-white colony screening (**C**). The transfection success of *Sf9* cells can be inspected with either light microscope (**B**), and due to YFP fluorescence, with UV-light microscope (**A**) with scale bars representing the magnification.

4.4 Recombinant protein expression in *S.sprugiferda*

The cell culturing was done using Spodoptera frugiperda 9 strain, infected with Autographa californica nucleopolyhedrovirus (AcMNPV). The cell cultures were prepared into the TNF-FH medium (Sigma), added with 5% fetal bovine serum, 5% inactivated fetal bovine serum, 10% symperonic® F68, 20 mM L-glutamine, 100 µg/ml penicillin-streptomycin. For protein expression cells were divided with the density of 50% of the 100/20 Corning cell culturing dish (Greiner). The cells were infected by inoculating 1:100 ratio of the virus titer into the culturing volume. The infected cell was incubated at +27°C and harvested after 72h (5min, 1000RPM, RT) and re-suspended into 137 mM NaCl, 2.7 mM KCl, eight mM Na2HPO4, 2 mM KH2PO4-buffer (PBS). Cells were homogenized in 10mM Tris (pH 7.8), 100 mM NaCl, 100 mM Glycine, 0.01 % Triton-X100-buffer, here on referred as affinity buffer. To prevent proteolytic degradation, protease inhibitor cocktail containing, mainly serine and cysteine proteases (Roche, product specifications not available), were used. The lysis was completed by using glass Teflon-piston homogenizer.

4.5 Protein purification

The lysate was clarified using centrifugation (20min, 20000 x g, +4°C), and filtered through 0,22 µ filter. The initial purification step was done using immobilized metal ion affinity chromatography. The supernatant was loaded into 5ml Histrap HP affinity column (GE Healthcare). Additionally, depending on the expression volume, the supernatant was loaded into 1.5x15cm econoflow gravity column (Biorad) loaded with 2.5 ml HisPur Ni-NTA-, or Talon Co2+ affinity resin (Thermo Fischer). The column was washed with 10 column volumes of washing buffer, containing 1x affinity buffer and 10 mM Imidazole (pH 7.8). The protein was eluted with three-step gradient elution ranging from 30mM-, 60mM- and 500mM imidazole in 1x affinity buffer. The typical elution profile varied 10ml with the peak fraction

of 2ml when using Äkta (GE healthcare), or Duoflow (Biorad) purification systems. The fractions were pooled and then concentrated to appropriate volume using 10000 MW Amicon Ultrafiltration devices (Merc). Additionally, the sample was desalted using a PD-10 desalting column (GE Healthcare).

The size exclusion chromatography was used for the polishing step. This was performed using Superdex 200 16/60 Hiload, or Superdex 200 10/300 increase gel filtration columns (GE healthcare), depending on the application. The peak fractions were pooled and, concentrated as described above. The protein concentration was determined using Nanodrop UV-spectrometer.

Sample fractions from the purification steps, cell lysis, soluble protein fractionation, insoluble protein fractionation, affinity purification, and gel filtration were collected for SDS-PAGE analysis. The samples was prepared into 0.05 % bromphenol blue, 2% SDS, 0.25 m DTT, 0.05 M Tris (pH 6.8), 2% glycerol- buffer and loaded onto 10% SDS-PAGE gel. The gel was run 200V, 50min. The gel was stained with Coomassie blue, and de-stained with 10% acetic acid, 10% ethanol-solution.

4.6 Protein deglycosylation

The glycosylation was performed using fusion protein, endoglycosidase H from *Streptomyces picatus*, and maltose binding protein (New England Biolabs). For analytical experiments (CD, analytical chromatography), 0.2 U/µl of Endo Hf was SEC purified sample fraction and incubated in 4℃ for 1 h to o/n. The fusion protein was removed using amylose resin affinity matrix (New England Biolabs) and then eluted with 1 CV of 10 mM D-maltose. For crystallization experiments, 0.1 U/µl of Endo Hf was added into the crystallization fraction, prior to plate set up.

4.7 Multi-angle light scattering (MALS)

The molecular mass and sample heterogeneity were determined using miniDAWN TREOS II multiangle light scattering detector (Wyatt technologies inc.). The system was coupled with a Shimadzu Prominence liquid chromatography system. The sample was applied to the Shimadzu SIL-20A autosampler unit. For size exclusion separation either s200 10/300- or, 5/150 increase columns (GE Healthcare) was used. The Prominence system configuration consisted of the SPD-20A photo-diode array detector, RF-20A fluorescence detector, and the RID-20A

refractive index detector. The system was equilibrated with either 10 mM HEPES- or Tris-based buffer until the laser- and the RI-detector had stabilized down the 10^{-3}-digits compared to the baseline. The obtained data were processed using ASTRA software (Wyatt technologies inc.)

4.8 Circular dichroism

The protein secondary structural elements and folding dynamics were analyzed using Circular dichroism. The instrument used was Chirascan CD spectrometer (Applied photophysics Ltd). The system was stabilized with constant nitrogen flow. The flow was adjusted to 0.3 dm3/min for the detector, and 0.1 0.3 dm3/min for the monochromator and the tungsten lamp. The sample was applied into 2mm thick quartz cuvette in the final concentration of 0.2 mg/ml. The system was zeroed with air measurement. The maximum spectrum coverage, was achieved, by blank measurement using 1:10 dilution of the affinity buffer. The single regular measurement was done scanning tree repetitions of the spectrum from 260 nm to 195 nm (1mM Tris pH 7.8, 10mM NaCl, 10 mM Glycine, 0.2 mM CaCl2) in 22℃. The melting temperature curve was determined using thermometer probe ranging from 22℃-95℃ with 0.5℃ degree ramp. The data was processed using Pro data viewer-, and Global3 software (Applied photophysics Ltd).

4.9 Quasi-elastic light scattering (QUELS)

The stability and solubility were analyzed using Dynapro Platereader II (Wyatt Instruments inc.) direct light scattering instrument. The system was pre-cooled to 10℃, and the laser was stabilized. The samples were applied into corning 386 DLS specific plate, with a final concentration of 0.05-0.5 mg/ml. The samples were analyzed by exposing the sample wells for 10 sec. for 12 image intervals. The laser signal was zeroed by inserting water and air blanks to first wells. The results were analyzed using the Dynamics software (Wyatt technologies inc.).

4.10 Radionuclide activity assay

The uncoupled 2-OG decarboxylation activity was determined with a modified prolyl-4-hydroxylase hydroxylation assay (Kivirikko & Prockop. 1972). The assay was done using purified protein fractions with IMAC. The assay was used to determine the time curve of the decarboxylation, the Km values of the cofactors and the total activity of the enzyme. Protein

concentration used in all of the measurements was 1µM. The 2-OG was labeled with 1.2µCi alpha-ketoglutaric acid, disodium salt (1-C14) (Perkin & Elmer). The cofactor solution (50µM FeSO2, 2mM L-ascorbic acid, 100µM, 1:100 catalase dilution, 50mM Tris pH 7.8, 2mg/ml BSA), was added to each reaction. The reaction was initiated by adding the 2-OG label and sealing the reaction tube. The released carbon dioxide from the carboxylation was captured into a Whatman paper that was pre-treated with Soluene-350. The typical reaction time was 30 min at 37°C. The oxygen Km measurements were determined under pO2-gradient. The gradient points were 21-, 16-, 8-, 4-,2- and 1%. The gradient was created using Ruskin InVivo400 hypoxia workstation system (Baker co.). The reaction was stopped with ice incubation, and the labeled carbon dioxide was released with 2:1 1M K2HPO4 pH 5.0. The reactions were incubated at room temperature for additional 30min with slow rocking. The Whatman papers were transferred into scintillation containers and soaked in toluene-2-methoxyethanol scintillation solution containing 1% polyphenylene oxide (PPO) and 0.003% 1,4-Bis(5-phenyl-2-oxazolyl)benzene (POPOP). The activity was measured with Tri-Carb 2900TR liquid scintillation counter (Perkin & Elmer).

4.11 Crystallization and X-ray crystallography

Protein crystallization experiments were carried out using the sitting drop method. The protein sample was pre-treated with endoglycosidase H – MBP fusion protein (New England Biolabs). The sample was briefly centrifuged and applied into a 96-well sitting drop iQ plates (TTP lab tech). The screens were applied using Mosquito LCP nano dispenser (TTP lab tech). The incubation and imaging were done using Rock imager 54- system (Formulatrix). The crystal inspection was done using the in-house remote software. Crystals were tested with the home source X-ray beam. The crystals were installed into an Oxford Cryostream 700 cryo unit. The X-ray generation instrument used was Bruker Microstar X8 Proteum. The system was integrated with Helios MX focusing optics and kappa goniometer. The acceleration voltage of the rotating Cu-anode was 20kV, producing fixed 1.5418 Å X-ray beam. The detector was Platinum 135 CCD detector.

4.12 Small-angle X-ray scattering (SAXS)

Small-angle X-ray experiments were carried out in two synchrotron locations. ESRF B29 beamline at Grenoble, France, and PETRA III P12 at Desy-Hamburg, Germany (Blanchet et al.

2015). The datasets were collected either with batch- or HPLC-mode. The Data collection was carried out by colleagues in Dr. Rajaram Venkatesan´s and Prof. Petri Kursula´s groups. The sample preparation consisted purifying the protein as described above. Dialysis was done using 3.5kDa mini tube completing three rounds with 10000:1 ratio of sample buffer (30mM HEPES (pH 7.5), 100mM NaCl, 100mM Glycine, and additionally 2mM CaCl2. 20µM FeSO4 was added into the samples just before shipping, to prevent auto-oxidation. The samples were shipped in sealed thermos system in 4°C. HPLC-SAXS measurement was done with superdex 200 2.3/300 analytical column. The sample concentrations in these datasets were 1.0 mg/ml (no FeSO4/CaCl2) 2.2 mg (no FeSO4/CaCl2), 1 mg/ml (FeSO4/CaCl2), for the ESRF dataset. The sample concentration for the PETRA III dataset was 0.5 mg/ml, 1mg/ml for the batch mode and 3.6 mg/ml for the HPLC mode

The datasets were processed using the ATSAS software, GNOM, and PRIMUS packages (Franke et al. 2017, Konarev et al. 2003, D. I. Svergun. 1992). Additionally, the HPLC sample data frames was decompressed with CHROMIXS (Panjkovich & Svergun. 2018). The molecular weight values were calculated using GNOM based on the Guinier approximation (MW I0) and the calculated molecular weight based on the scattering of a BSA molecule (MW calc). Two types of de novo molecular modeling were done. Ab initio modeling was done using two ATSAS software, packages, DAMMIN (D. I. Svergun. 1999) and GASBOR (Petoukhov et al. 2012). The output solutions were averaged, and surface minimized using DAMAVER (Volkov & Svergun. 2003). The output was aligned for evaluation using SUBCOMB (Kozin & Svergun. 2001) The rigid body modeling was done using BUNCH (Petoukhov & Svergun. 2005). In addition, the crystal structures used in rigid body modeling was evaluated using CRYSOL (D. Svergun et al. 1995). The structure graphical interpretation and representation was done using pyMOL (Schrödinger. 2018).

5. RESULTS AND DISCUSSION

5.1 Comprehensive protein bioinformatics

The contradiction regarding P4HTM is that it has more collagen P4H structural homology than to HIF-P4Hs. The ability to hydroxylate HIFα ODDD peptides, suggest that P4HTM contributes to the HIF-pathway in some levels (Koivunen *et al.* 2007a, Laitala *et al.* 2012, Leinonen *et al.* 2016). Since the identification of P4HTM, many P4H crystal structures have been available. Structural prediction algorithms and sequence alignment servers were used to asses the homology among P4H-enzymes. The homologous sequences and the structural modeling was done using structure prediction programs and multiple sequences — the modeling based on the available crystal structure coordinates. Thus the 2-OG dioxygenase domain is expected to be covered, yet the transmembrane region and the vicinity of the N-terminus are not likely to appear. The results were then used to aid the construct design, and to confirm the assumed characteristic differences of P4HTM.

5.1.1 Multiple sequence alignment

The sequences selected for the multiple alignments were chosen with two primary conditions. The distribution has to contain different species and that the sequence appears in raw sequence check. In order to compare the relation to HIF- or C-P4Hs, human HIF-P4H1, -2, -3, C-P4H-αI was selected. Other chosen homologous were Cr-P4H (*C.reinhadrtii*), Ba-P4H (*Bacillus anthwaxis*), PBCV-1 (*Paramedicum bursaria* chorella virus-1). The primary alignment and the sequence manual editing was done as described above (see materials and methods). In parallel for structure prediction purposes, the template searchin engines resulted in multiple outputs linking to Ca^{2+} binding proteins, such as calmodulins.

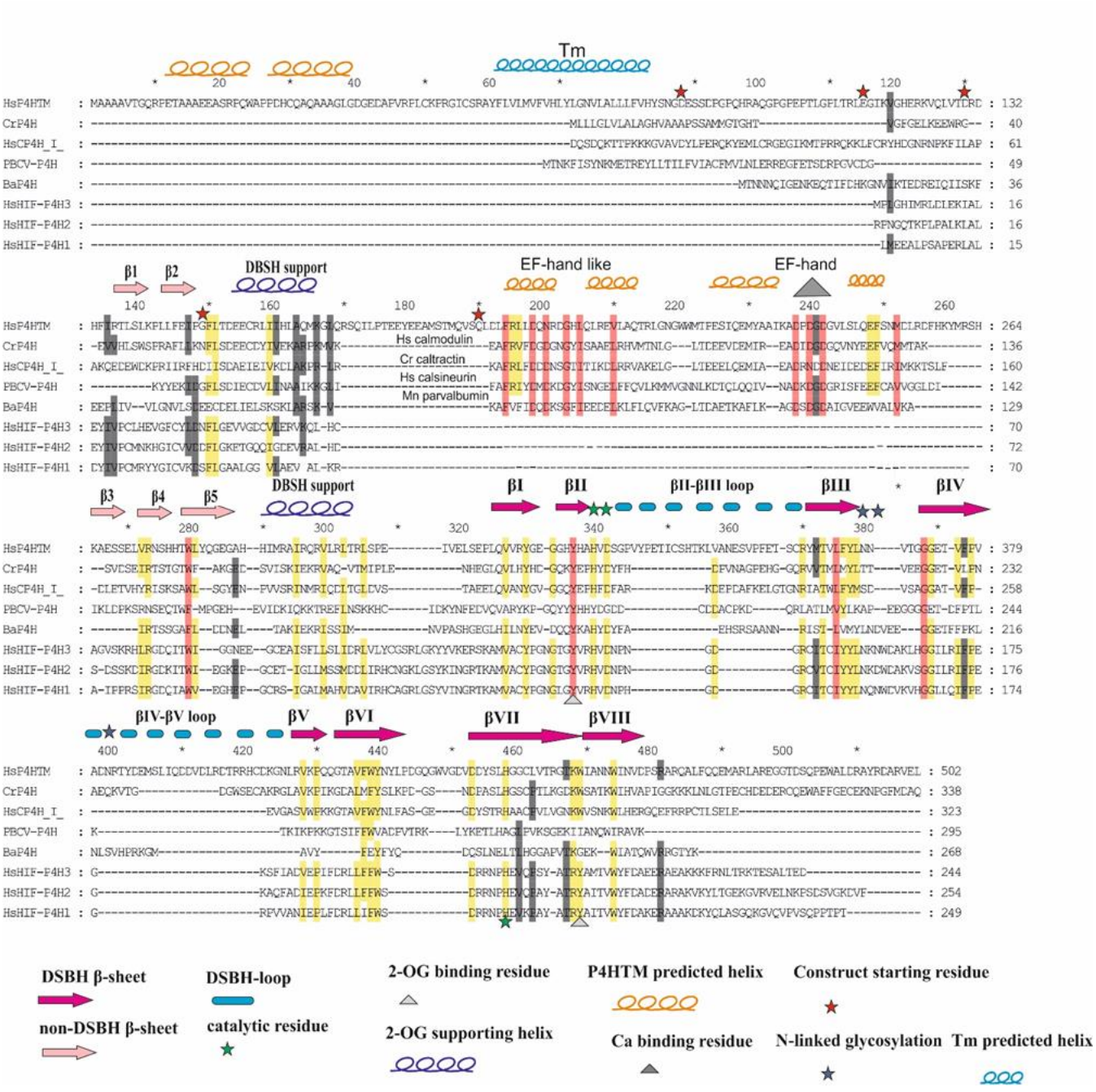


Fig.7 Multiple sequence alignment of full-length P4HTM residues 1-502. The Inputs on the left column are displayed in two groups, C-P4Hs (four sequences) and HIF-P4H (three sequences). The N-terminal part of P4HTM resembles the transmembrane domain. The calcium binding motifs were searched using the inputs from the structure prediction results. The alignment of these proteins start at aa. 180 and ending at the end of the row. The protein names from top to below are human calmodulin, *C.renhardtii* caltractin, human calcineurin, and *M.merlangus* parvalbumin. The DSBH-fold is displayed as βI-βIII and their locations are based on the Cr-P4H structure in parallel with the secondary structure prediction output from phyre2, SWISS-

model, HMMTOP and DAS (Koski *et al.* 2007). The P4HTM constructs present the starting points of the current TmΔ88-502 construct and the shorter ones (see materials and methods). The detailed description of the alignment is displayed at the bottom.

The feature in P4HTM sequence was discovered by Koivunen and collegues (2007) and this alignment confirms it. The β-loop before the DBSH-fold contained a motif with residues that are present in calcium binding proteins. More detailed view showed that the motifs are conserved in P4HTM. In the (Koivunen *et al.* 2007) publication the region was shown to align with C-P4Hs and HIF-P4Hs. However, the new alignment shows that the whole EF-hand (Kretsinger & Nockolds. 1973) motifs are conserved. The EF-hand region aligned well with homologous sequences confirming that P4HTM likely binds to calcium and potentially be necessary to the biological function. The transmembrane sequence input was put through the membrane prediction engine, which resulted in a single helix through the membrane. The structure prediction servers identified the locations of the predicted transmembrane domain and the N-terminal part in the cytosol. The lumen part was covered mostly with the same secondary features found in Cr-P4H. Thus the coordinates 2JIG was used to present the possible secondary structure of the 2-OG dioxygenase domain (**Fig.7)**.

The sequence alignment suggests that the composition and order of the 2-OG dioxygenase domain resemble most likely the DSDH-fold. In almost all sequences the catalytic and co-factor binding residues are conserved. However, the exact lengths and orientations of the β-sheets are likely to be different. In almost all sequences the catalytic and co-factor binding residues are conserved. However, the exact lengths and orientations of the β-sheets might vary. The alignment also confirms the modeling results, suggesting that the P4HTM has more shared features with C-P4Hs than HIF-P4Hs (fig.7. The conserved group is highlighted, but the residue is different.

5.1.2 Tertiary structure predictions

Majority of the sequence residue hits, coverage and the model confidence appeared in the 2-OG dioxygenase domain part. Both programs yielded in all possible cycles, three distinct classes that appeared in the results repetitively. The highest sequence alignment coverage and identity came from Cr-P4H inputs. The Cr-P4H crystal structures, PDB entries 3GZE, 2JIG,

2JIJ (Koski et al. 2009a, Koski et al. 2007), appeared in all cycles with the sequence coverage up to approximately 40%, including the catalytic residues. The second most appearing class identified from the Bacillus anthraxis P4Hs (Ba-P4H), PDB entries 3ITQ, 5HV0 (Culpepper et al. 2010, Schnicker & Dey. 2016), and Paramedicum bursaria chorella virus-1 P4Hs (PBCV-P4H), PDB entries 5C5U, 5C5T (Longbotham et al. 2015). The sequence coverage of the crystal structures was approximately 35%. The third appearing group covering the 2-OG dioxygenase domain was the HIF-P4H crystal structures, PDB entries 3OUJ, 2G19 (McDonough et al. 2006, Rosen et al. 2010). Some less conserved results regarding the 2-OG domain came from Lysyl-hydroxylases (Guo et al. 2018) and other prolyl-4-hydroxylases from the lower organism P4Hs. It was clear that based on the modeling results, the majority of the models based on collagen P4Hs and fewer HIF-P4Hs. All models with over 95% confidence were then annealed, resulting in the final prediction structure. The final model in Phyre2 covered 49% of the P4HTM sequence (**Fig.8**). By looking on the available crystal structures, the P4HTM model resembles the DSBH-fold

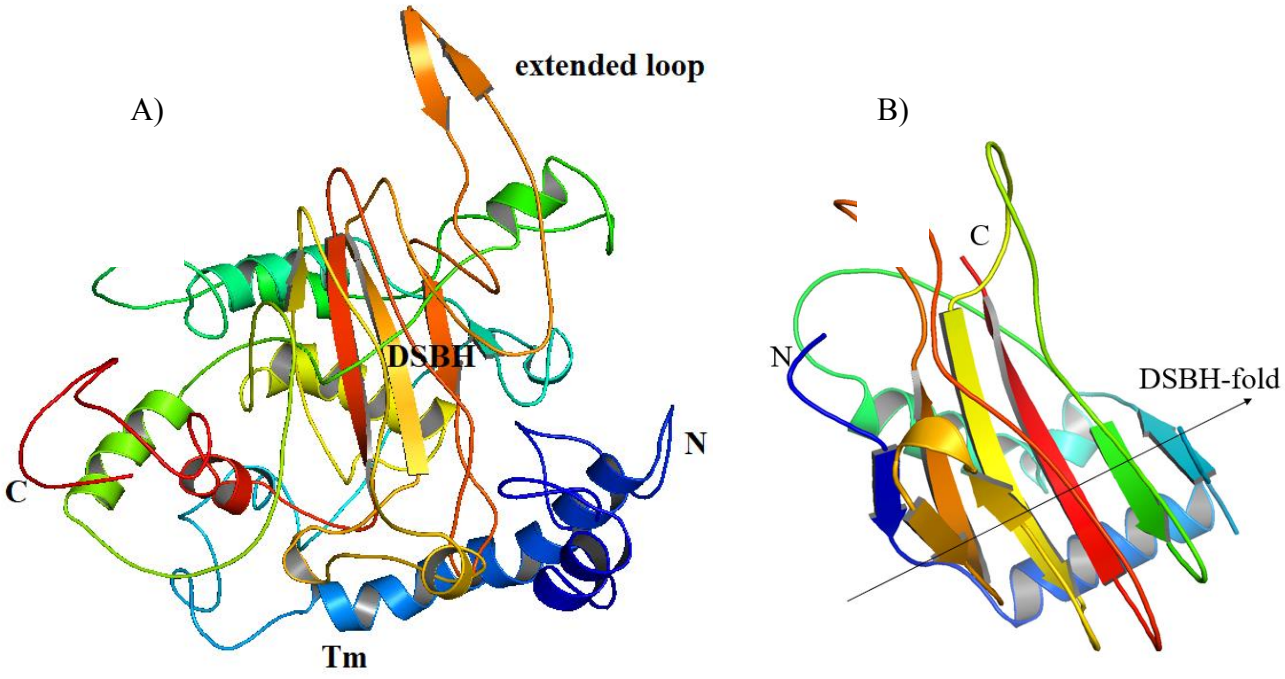


Fig.8 Models of P4HTM. The Phyre2 annealing resulted in a model with 49% coverage with 95% coverage (A). The model features the DSBH fold and the transmembrane helix. The loops and the possible EF hand is too disordered to distinguish. However the βII-βIII loop seems to appear in the rear. The presence of the fold can also be seen in some more coarsed models (B).

Other than the 2-OG dioxygenase domain, the EF-/EF-hand like loop structure was well covered and usable for modeling. Many of the outputs suggested that the P4HTM EF-hand region is similar to calmodulins, caltractins, calcineurins, and other calcium-binding domains, but not so many of these models were based on human proteins (Schuermann et al. 2010, Ye et al. 2013). Many of the outputs came out as merely calcium-binding proteins especially from animals with approximately < 20% sequence of the full-length P4HTM

The sequence similarity of P4HTM with other P4H enzymes is the highest in the C-terminal region suggesting that the C-terminal part of the enzyme consists mainly out of DSBH-fold. In almost all sequences the catalytic and co-factor binding residues are conserved. The alignment also confirms the modeling results, suggesting that the P4HTM has more shared features with C-P4Hs than HIF-P4Hs, based on the model input list, which was used for the modeling. The fit of the annealing was validated manually, by visualizing the available crystal structures. Because the modeling software suggested Cr-P4H as the closest homolog, the crystal structure (PDB entry 2JIG) served as a reference of aligning the DSBH-fold respect to P4HTM model (see fig.8a). The Cr-P4H structure was used as a reference to compare the secondary structure prediction of P4HTM to the coordinates. The visualization showed that the confidence given in the program was true, at least concerning the DSBH core fold. According to the sequence alignment, the catalytic residues in P4HTM is H328, D330, and H374. The 2-OG binding residue is lysine (K345), as in all C-P4H enzymes, whereas in HIF-P4Hs the residue is arginine (see fig.7). Unique shared feature with Cr-P4H is the conserved cysteine residues C404-C444. The corresponding cysteine Cr-P4H residues form a disulfide (Koski *et al.* 2007). Two of N-glycosylated asparagines locates between the βIII-βIV sheets and one at before the βIV-βV loop.

5.2 Protein purification optimization of the TmΔ88-502 construct

The recombinant truncated TmΔ88-502 construct was described by Koivunen and colleagues (2007) has proven to be soluble *in vitro*. The full-length P4HTM was seen to be proteolytically cleaved after the transmembrane from this residue. For crystallization purposes, the purification process was optimized, in order to maximize the quantity of pure protein. The purification was done as described above (see materials and methods), which is based on the (Koivunen *et al.*

2007) publication. The typical purification final yields obtained from the TmΔ88-502 purification were low (**Table.**I). The protein expresses well in insect cells. The protein appears to elute as a monomer based on the retention volume, which appears to elute where BSA standard elute (~15ml in s200 10/300 column). However, during the purification, visually estimated >60% of protein is lost at final purification stage >90% pure (**Fig.9**). Since the decrease is gradual, it would suggest that the protein tends to aggregate and eventually precipitate. An ordinary protein purification protocol requires concentrating, especially before high-performance liquid chromatography (HPLC).

Table.I The typical purification results obtained from the TmΔ88-502 purification.

Expression volume (ml)	Final protein concentration (mg/ml)	Purity (%)	Final volume (µl)	Yield (mg)	Yield (mg/l)
500	2.5	>90	50.0	0.125	0.25

The structure prediction suggests that the region between the 2-OG dioxygenase fold and the transmembrane region may contain distorted random coil region. The disordered regions might cause flexibility to the protein. The most significant feature regarding the solubility comes from the three possible N-glycosylated residues. Authors previous attempts to express the protein in *E.coli* revealed that it is challenging to obtain folded soluble protein in *E.coli* system with the use of detergents and chaperone co-expression (data not shown). The absence of glycosylated residues causes the majority of the protein to separate in the insoluble fraction, which suggests that the glycosylation is crucial for the solubility. The expression in *S.frugiperda* enables sufficient expression and yielding more soluble protein. However, the glycosylation in the ER is a complicated process, and it is much limited in *insecta* than it is in higher *eukaryotes* (Shi & Jarvis. 2007). Therefore, the glycosylation efficiency may vary on every protein batch and increase the insoluble fraction.

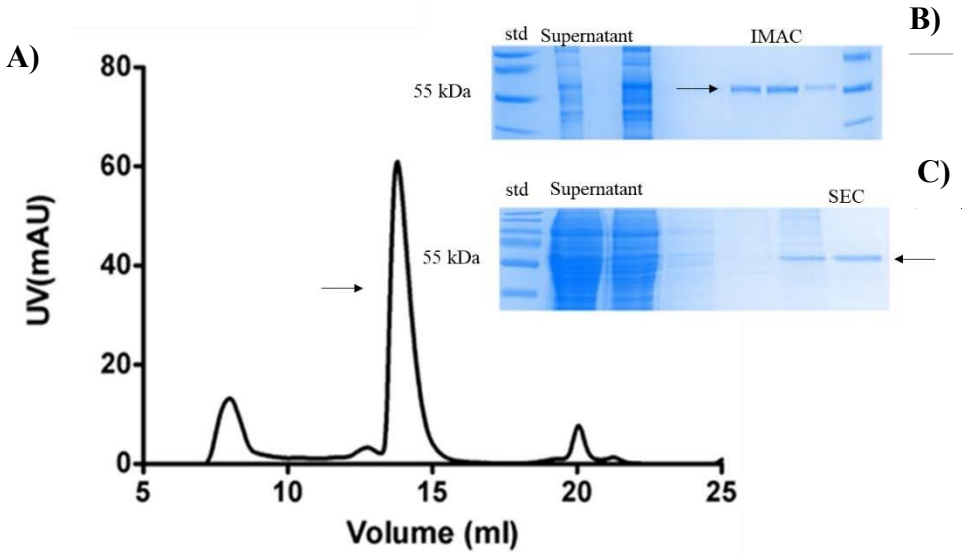


Fig.9 Typical result of TmΔ88-502 purification. The retention volume of the soluble peak corresponds to a ~55 kDa protein (A). The protein expresses well in insect cells, as seen in the supernatant. However, the yil during the purification, yield decreases. The IMAC purified fraction displayed in the upper SDS-PAGE (B) and the SEC purified fractions in the lower gel (C).

In order to improve the purification yield, quasi-elastic light scattering experiment was done to screen various buffering conditions for an HPLC-purified fraction (see materials and methods). The screening conditions consisted mainly of GOOD buffer solutions (Good *et al.* 1966). A variety of additives such as salt, glycine, $CaCl_2$, and DTT was tested. The screening conditions were in an ambient temperature close to the experimental temperature range +4 - +10 °C (**Fig.10**). The QELS measurement is an excellent indicator of any large aggregated species that are present in the sample matrix, but not an accurate technique to determine the molecular mass of the particle. The more ideal parameters for DLS is the radius of gyration and the polydispersity of the particle (Stetefeld *et al.* 2016).

The DLS spectra showed that the protein is very insoluble and prone to aggregation. The polydispersity and most conditions suggest that the sample is heterogeneous, meaning that the sample is a mixture of soluble and insoluble species. In order to detect any improvement and determine the most suitable buffer system for the protein, the interfering signals were cut. The dimensions of P4HTM monomer was approximated to 5 nm (La Verde *et al.* 2017, Pronk *et al.* 2014), based on the calculated molecular weight. The effect of additives could stabilize the protein on a microscopical scale. Generally, the same distribution was seen in samples without

the additives, thus indicating no improvement. The redox state of the protein did not seem to be affected by the reducing agent. Usually, the reducing agent helps the stability of assumed disulfide bonds and preventing unwanted isomerization, by keeping the cysteines reduced (Gräslund *et al.* 2008, O'Shaughnessy & Doyle. 2011). The literature shows that the human C-P4H truncated variants yield and purity increased by the addition of chelating agents such as EDTA. Addition of EDTA prevents oxidization of disulfides, causing disulfide isomerization and misplaced disulfides (Murthy. 2018). The reason being that atmospheric oxygen might oxidize thiol groups and inducing disulfide. The protein was expressed in *Sf9* which means that the protein should have the ability to catalyze proper possible disulfide formation in the ER.

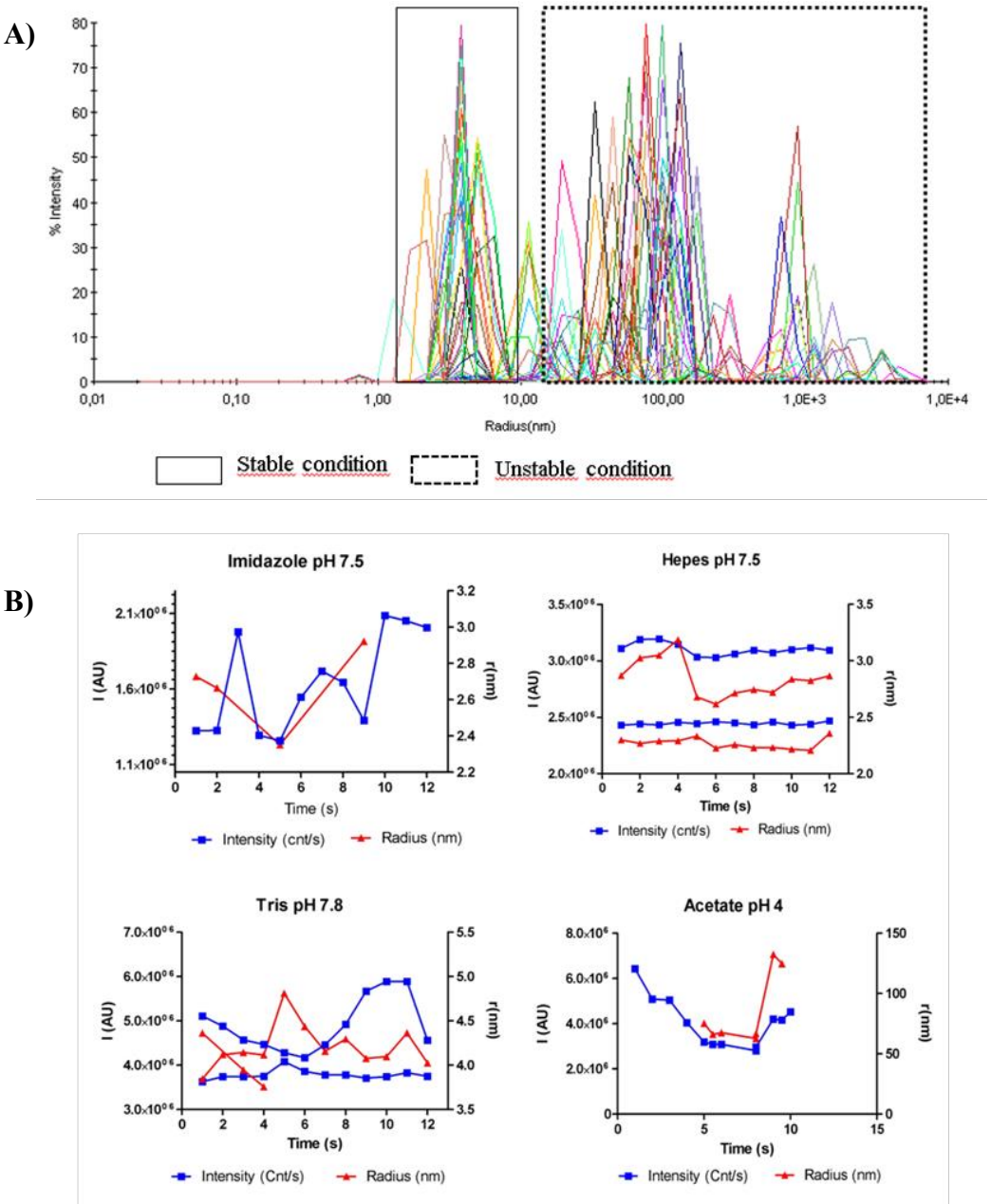


Fig.10 The overview of the QELS-spectra (A), shows that majority measurements the sample matrix appears as heterogeneous and multidisperce. The mass distribution and the particle

radius of gyration in the particles are broadly distributed indicating that the protein is not very stable. The linearized graphs of the particle radius show that neural pH has the correct size (B). In an acidic condition (control) such as in acetate, the protein starts by forming aggregates, causing the radius to increase. The sample concentration in this measurement was 0.1mg/ml.

The protein associated with the ER membrane most likely means that the protein has some hydrophobic parts in the fold. Even though the TmΔ88-502 construct does not contain the transmembrane domain; the protein might associate with any membranes. Another aspect as mentioned above is that the efficiency of the glycosylation may vary. The insoluble/soluble fraction ratio varied in different purifications, thus making difficult to establish consistent purification results. Overall it the results show that it was challenging to purify >90% pure fractions using this construct.

5.3 Oligomeric state and molecular mass determination

The oligomeric state of the TmΔ88-502 construct oligomeric state was determined based on the elution profiles of size exclusion chromatography and SEC-MALS (**Fig.11**). As discussed earlier the main soluble peak retention volume suggest that the protein elutes in the same range as BSA which means that the protein appears as a monomer *in vitro* rather than homodimer (discussed in the literature review). In the previous sections, the SDS-PAGE of SEC purified fraction from glycan containing protein has approximately ~55 kDa mass and the deglycosylated protein is approximate ~50 kDa. The aggregated peak is not observable in SDS-PAGE. The multiangle scattering experiment was done in the buffering condition as presented in the materials and methods section. The possible solubilizing effect of the use of EDTA was tested by adding 1mM EDTA to the sample. The mass distribution analysis suggests that the construct is a monomer *in vitro* and appearing as a heterogeneous mixture.

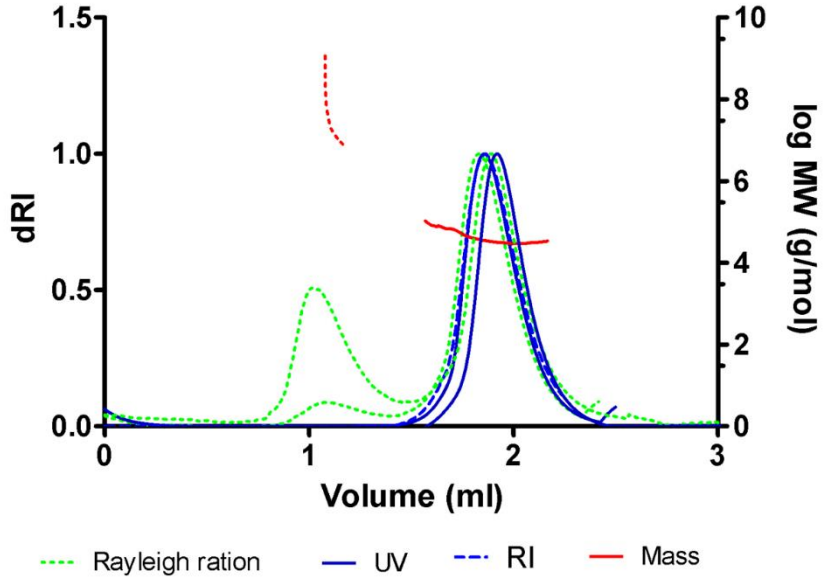


Fig.11 Determinating the oligomeric state of the TmΔ88-502 construct. The chromatogram displays the mass distribution over the peak. The molecular mass was computed based on the differential refractive index corresponding to roughly 50 kDa peak. The peak is non-symmetrical indicating that the peak contains additional species, which might be caused by cleavage of the protein, or the heterogeneity in the glycosylation. Dimerization is not likely to occur. The EDTA addition does not seem likely affect the solubility since the insoluble peak is higher (Rayleigh ratio, green dots).

The molecular mass estimate determined with SEC-MALS show that the mass distribution range from 34-54kDa, which is acceptable within the limitation of the measurement. The results are consistent with previous measurements done on the same construct (Unpublished data). The main peak seems to contain a mixture of species, but none of them suggest that the protein appears as a dimer as described previously (Koivunen et al. 2007). The higher molecular weight as respect to the calculated mass can be explained by the N-glycosylation (Mehta *et al.* 2016). The lower molecular mass could be caused by the fact that the N-terminal part is very flexible resulting inaccurate measurement of the particle. It is also possible that some degradation might occur, but it is not likely that the monomer would decompose into two fragments.

5.4 Kinetic measurements

Radionuclide-labeled 2-OG uncoupled activity assays were done to determine the K_m values for oxygen and iron. The oxygen measurements were done ranging the oxygen partial pressure from 21% to 1%. The reaction time for the condition was 30 min. The iron measurements were done as described in the materials and methods section. The reaction time for the iron reaction was 5 min. The time was determined to perform time-curve measurements with fixed Fe^{2+} concentration. The protein concentration used in these reactions were 1µM, and the 2-OG concentration was 100mM. The slope of the activity was determined by plotting the values of the decayed CO_2 as a function of concentration, and the fit was calculated based on the sum of squares (**Fig.12**). The measurement should be considered as an indirect measure of activity since only the measured parameter is the decarboxylation of 2-OG (Kivirikko *et al.* 1989). As seen elsewhere, the K_m values of the co-factors do not change in the presence of the substrate (see review (Myllyharju. 2013). The reaction is still valid of determining the kinetics of enzyme and the co-factors. The oxygen measurement shows that the P4HTM has much lower K_m value as respect to HIF-P4H2 (**Table.II**).

Table.II The K_m values for O_2 and Fe^{2+}. The values were determined in three individual measurements and given as mean standard deviation (SD). For reference values see (Mcneill et al. 2002 (a) Hirsilä et al. 2003 (b), Hirsilä et al. 2006 (c))

Enzyme	Km (µM) Fe^{2+}	Km (µM) O_2
TmΔ88-502	0,98 +- 0,085	85 +- 12
C-P4H α1	2[a]	40[b]
HIF-P4H1	0,03[c]	230[b]
HIF-P4H2	0,03[c]	250[b]
HIF-P4H3	0,1[c]	230[b]

The P4HTM K_m for oxygen differs from HIF-P4Hs, where the K_m range for ODDD peptides is 230-250µM. The K_m decreases down to 100 µM when the peptide elongates with (Ehrismann *et al.* 2007, Hirsila *et al.* 2003, Koivunen *et al.* 2006). The lower K_m value respect to HIF-P4Hs, suggest that P4HTM is not primarily oxygen sensor, but the contribution to hypoxia is possible. The oxygen K_m values for C-P4H 40µM, are lower than HIF-P4Hs. The K_m value for iron is 2µM (Myllyharju & Kivirikko. 1997). The K_m iron value for P4HTM indicates that the reaction rate is high, and possibly with binding with high affinity (Dao *et al.* 2009). Since the K_m value

of iron is low, the binding affinity would provide more information about the binding of Fe^{2+} and also Ca^{2+}. The actual binding of Fe^{2+} and O_2 could be determined more accurately with other techniques such as isothermal titration calorimetry (ITC), or surface plasmon resonance (SPR).

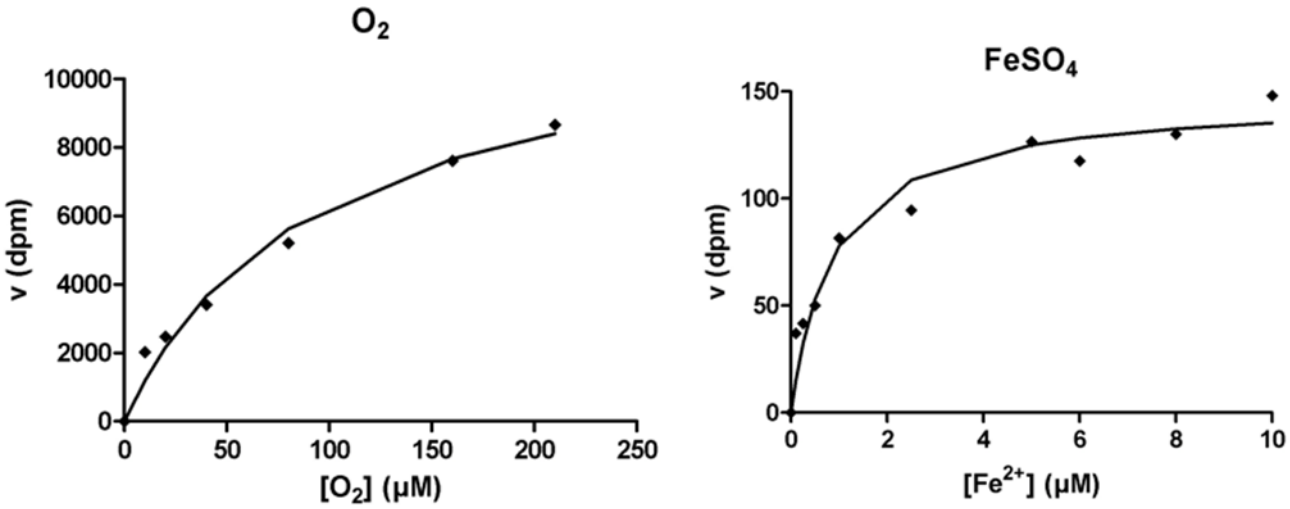


Fig.12 Km value determination for Fe^{2+} and O_2. The Km values were determined from three independent measurements and reported as mean standard deviation

5.5 Effect of metal ions

In addition to the Fe^{2+} coordination in the active site, most likely the Ca^{2+}-cation binds into the EF-/EF-hand like motif. As described in the bioinformatics section, the domain is located before the DSBH-fold and based on the structural homology it replaces the β4-β5-loop in Cr-P4H and β3-β4 loop in HIF-P4H. The loop could act similarly as the βII-βIII-loop which is most likely responsible for the substrate binding and specificity. The affect of metal ion absence is likely to affect the solubility and the structure. The protein was therefore purified without metal ions, and CD-spectra was measured (**Fig.13**).

The CD curves of metal-containing samples show more ordered secondary structure; the CD curves was deconvoluted to secondary structure ratio using CDNN (Greenfield. 2006) The the sample not containing metals showed the slightly disordered curve which is seen in the secondary structure variation (**Table.III)**

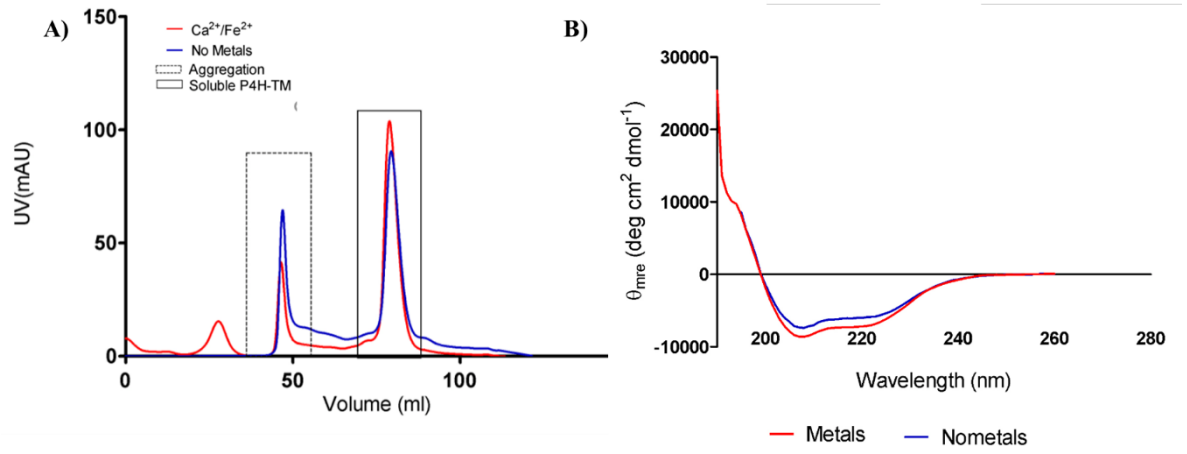


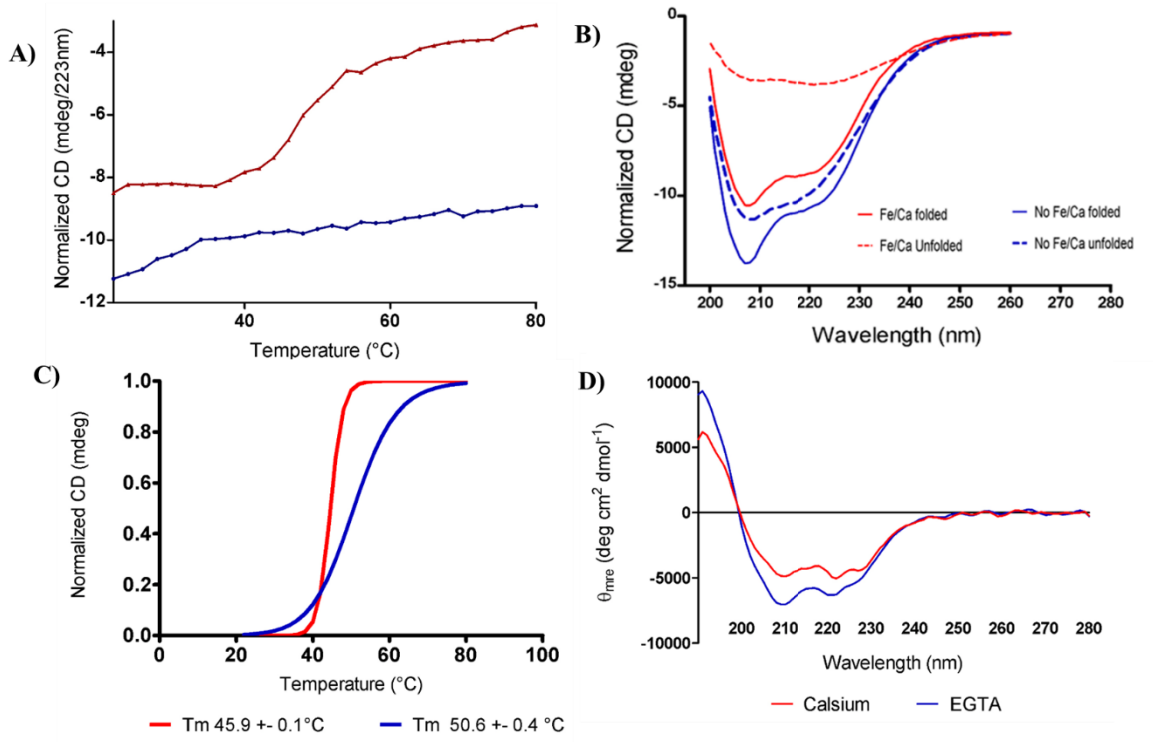
Fig.13 The size-exclusion chromatogram shows that the ratio of insoluble fraction respect to soluble fraction increases with the absence of metal ions (A). The CD-spectra suggest that the there is minimal difference in the secondary structure (B). The CD measurement was done with 0.2 mg/ml protein concentration.

Table.III CD-spectra deconvolution.

	no-metals				metals			
Helix	22 %	20 %	21 %	19 %	27 %	25 %	24 %	26 %
Antiparallel	19 %	14 %	12 %	14 %	18 %	16 %	12 %	10 %
Parallel	12 %	13 %	14 %	13 %	9 %	11 %	11 %	12 %
Beta-Turn	19 %	20 %	19 %	20 %	19 %	19 %	19 %	18 %
Rndm. Coil	39 %	43 %	45 %	45 %	32 %	35 %	38 %	39 %
Total Sum	111 %	110 %	110 %	111 %	105 %	105 %	104 %	105 %
	195-260 nm	200-260 nm	205-260 nm	210-260 nm	190-260 nm	195-260 nm	200-260 nm	205-260 nm

The possible stabilizing effect of the metal ions was tested with CD melting curves. The curves show that the metal-containing curve has a clear transition between the folded and unfolded state. However, when the enzyme is not bound to metals, the protein aggregates rapidly and do not show a clear transition. The effect of calsium In order to determine whether iron or calcium is more important to the folding stability, CD measurement was done separately for calcium conditions. The measurement was performed using EGTA as a calcium chelator, which inhibits Ca^{2+}-binding. The results show that there is contradicting difference in the CD curves. In the presence of calcium, the secondary structure shows to be more disordered than without the calcium. The data suggest that the coordination of the Fe^{2+}-cation is more critical for the protein stability *in vitro*. However, the absence of calcium may not affect the uncoupled decarboxylation activity. Also, the experimental setup cannot exclude the possibility of already

bound calcium. Further experiments require more knowledge of the binding kinetics and thermodynamics of the metal ions.



(**Fig.14**). The melting temperature (Tm) was found to be 42.8 **°C** ± 0.1 for the metal containing sample and 50.4 **°C** ± 0.4 applying the single-transition state model (A). The melting curves show that in the non-metal condition the protein is likely already aggregated as the temperature increases (A). The relative CD curve show also that there is no shift between unfolded and folded state (B).

5.6 Crystallization and diffraction tests

The crystallization experiments proceeded first with the available modified commercial screens. Some conditions were optimized mainly containing 2-(N-morpholino)ethanesulfonic acid (MES), (3-(N-morpholino)propanesulfonic acid) (MOPS) as a buffering component. The main precipitants in these conditions were polyethylene glycol (PEG), tert-butanol, and a mixture of chaotropic and kosmotropic salts such as $(NH_4)_2SO_4$ and KCl. Few promising crystals formed in conditions with the high amount of precipitant such as 100mM acetic acid pH 4.5, 1M $(NH_4)_2SO_4$, 1M KCl. The crystals were tested with the home-source X-ray beam, which proved to be salt crystals (Data not shown). The protein concentration used did not

exceed over 4 mg/ml, which is not ideal for crystallization. Taking to account that the protein is likely very insoluble, the limiting crystallization concentration might not be high. Higher yields and possible more efficient de-glycosylation protocol is required to improve these experiments.

5.7 Low-resolution P4HTM structure in solution

5.1.1 Data collection and processing

Due to difficulties of obtaining crystals, small-angle X-ray scattering was used to determine low-resolution structure in two conditions. The samples were purified and prepared without $FeSO_4$ and $CaCl_2$ and an as a comparison with the metals. The samples were measured as a batch sample, which uses a quartz capillary to place the solution in front of the beam. The second type of sample was an HPLC-coupled system with the beam (Malaby *et al.* 2015). The data sets were collected in two beamlines. The batch sample datasets were collected in ESRF BioSAXS BM29-beamline in Grenoble. Four samples were measured in the buffer containing $FeSO_4$ and $CaCl_2$ or without them. The sample concentrations in the non-metal condition were 1 mg/ml and 2.2 mg/ml. The normal condition sample concentrations were 1 mg/ml and 3 mg/ml. The batch mode measurement has a disadvantage that any possible intermolecular contacts cause an error to the measurements because all particles exposed to the photons, and the scattering curve is a differential average of the pattern. In case fraction of the sample is aggregated, it will cause an error in the particle dimension determination. Therefore, larger aggregates will positively increase the dimension and mass of the particle. The concentration determination is also crucial to the measurement (Skou *et al.* 2014). The negative effect of the aggregation was eliminated using HPLC-SAXS measurement, which allowed to collect data only from the soluble peak. The HPLC-SAXS dataset was collected at PETRA III P12-beamline in Hamburg. The sample was sent in containing the metal ions in the buffering conditions with a concentration of 3.6 mg/ml. The retention of the peak was similar as in the purification experiments. Surprisingly, there was no insoluble peak observed during the separation. The soluble peak was continuously exposed, and the baseline was determined to from zero mobile phase, i.e., buffer.

The scattering curves for the metal comparison measurement showed that there is some interparticle interaction difference in the smaller angles, causing positive repulsion. The non-metal samples seem to have some positive curvature, which is usually an indication of various species in the sample. Therefore, part of the sample is likely aggregates. The metal sample scattering curves show that there much less positive curvature. The concentration difference can is observed well. The sample containing cations had much less these effects. The shape of the scattering curve in the in the higher angles (1-3 nm^{-1}) suggests that the protein is globular and possibly elongated.

The Guinier regions for both samples showed linearity and indicating that the data is usable, for modeling purposes The Kratky plot and Porod plot show that the protein is in a folded state and is likely a globular monomeric particle. The radius of hydration for the metal-containing samples were 3.68 ± 0.01 (I_0 57, 1mg/ml) and 4.25 ± 0.28 (I_0 82, 3 mg/ml). The non-metal sample radius of gyrations were 3.67 ± 0.09 (I_0 58, 1 mg/ml) and 3.9 ± 0.11 (I_0 66, 2.2 mg/ml. The linear part of the Guinier region and the error in the R_g suggest that the sample might behave better in lower concentration. The based on the intensity and the particle volume the molecular weight the molecular weight was 95-130 kDa for the metal-containing the sample. The molecular weights for the non-metal sample were 89-110kDa. Due to the large range the scattering curves where merged which represent as a mean of the molecular mass. The molecular weight estimates suggest that in both conditions the protein is a dimer. However, considering that there are most likely interparticle interactions increasing the structure factor, the values are caused by the aggregation artifacts. The metal-containing sample Dmax ranges from 15 to 18 nm and the non-metal containing sample from 15 to 20nm. The measurement is done with the HPLC-mode, and another batch sample as a reference showed that the protein is a monomer. The Rg range 3.38 ± 0.02 (I_0 349 3,6 mg/ml) is agreement with the metal/non-metal samples. The molecular weight was calculated based on the BSA standard scattering to range from 49kDa to 59 kDa with a Dmax of 12 nm. The ESRF dataset is displayed in (**Fig.15**). The PETRA III datasets are displayed in (**Fig.16**). Previously the full-length P4HTM and the TmΔ88-502 variant was described as a homodimer *in vitro* (Koivunen *et al.* 2007a). Based on the SAXS measurements together with MALS measurements, the oligomeric state of this construct is a monomer. The results of the SAXS measurements are summarized in (**Table.II**).

Table.III SAXS data results. The metal comparison values are presented from merged scattering curves from two concentrations. The HPLC-sample values are from the single scattering curve obtained from the soluble chromatography peak.

Sample	Rg (nm)	I0	Dmax (nm)	Vp (nm3)	MWest (kDa)	MWcal (kDa)	$\chi 2_{\text{Dummin}}$	$\chi 2_{\text{Gasbor}}$	Datapoints	processed region	processing software	Photon source
Fe/Calsium HPLC	3,34 ± 0,02	348,33 ± 0,46	12	129,21	88	48	1,1	2,2	2634	23-853	GNOM	PETRA III P12
No Fe/Calsium	3,74 ± 0,05	59 ± 0,33	18	171,9	111	104	0,91	1,12	1043	36-432	GNOM	ESRF BM29
Fe/Calsium	4,2 ± 0,15	83 ± 0,17	19	172,90	124	110	0,92	1,05	1043	18-407	GNOM	ESRF BM29

Experimentally the results could be improved increasing the number of measurements and with broader concentration series. The HPLC coupled measurement contains more data points improving the Guinier region linearity. The size exclusion in real time excludes the possibility of interfering aggregates. Therefore, the data collection with the PETRA III beamline was more successful experimentally. The data parameters seem to indicate that there is no difference whether the metal ions affect the enzyme stability. However, since the scattering curve indicates aggregation, the metals may cause the fold the become more flexible thus inducing aggregation.

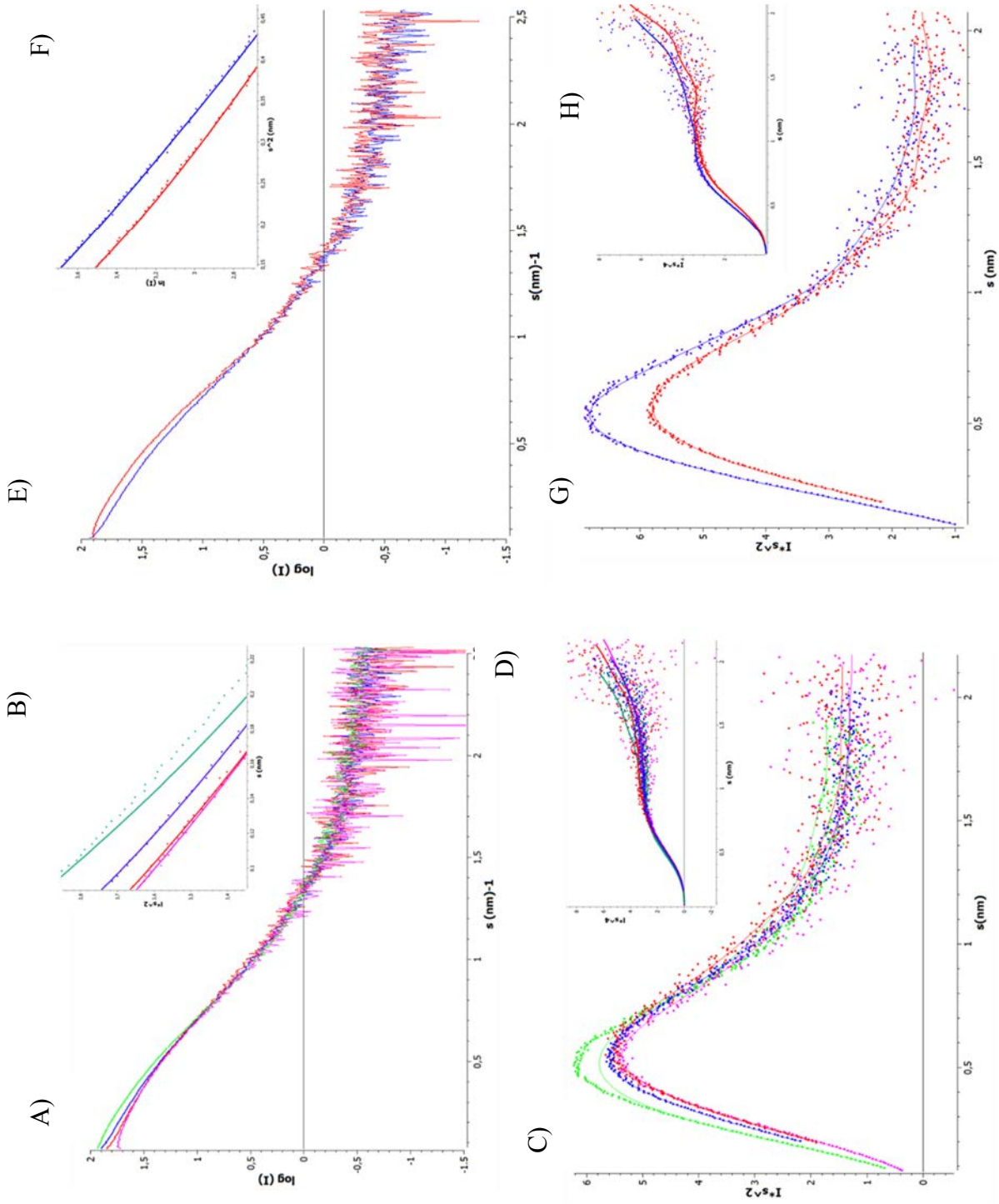


Fig.15 TmΔ88-502 construct measured in ESRF intwo conditions. A solution containing 20µM FeSO4 and 2mM $CaCl_2$. The sample buffer composition, were 10mM HEPES pH 7.5, 100mM NaCl, 100mM Glycine. The sample concentration containing the metals were 1 mg/ml and 3 mg/ml (magenta/green). The sample concentrations for the non-metal conditions were 1 mg/ml and 2.2 mg/ml (red/blue). The data were processed separately for each sample Scattering curve (A), Guinier plot (B), Kratky plot (C) and porod plot (D). The datasets were merged and

processed the same way respectively, scattering curve merged (E), Guinier plot merged (F), Kratky plot merged (G), porod plot merged (H)

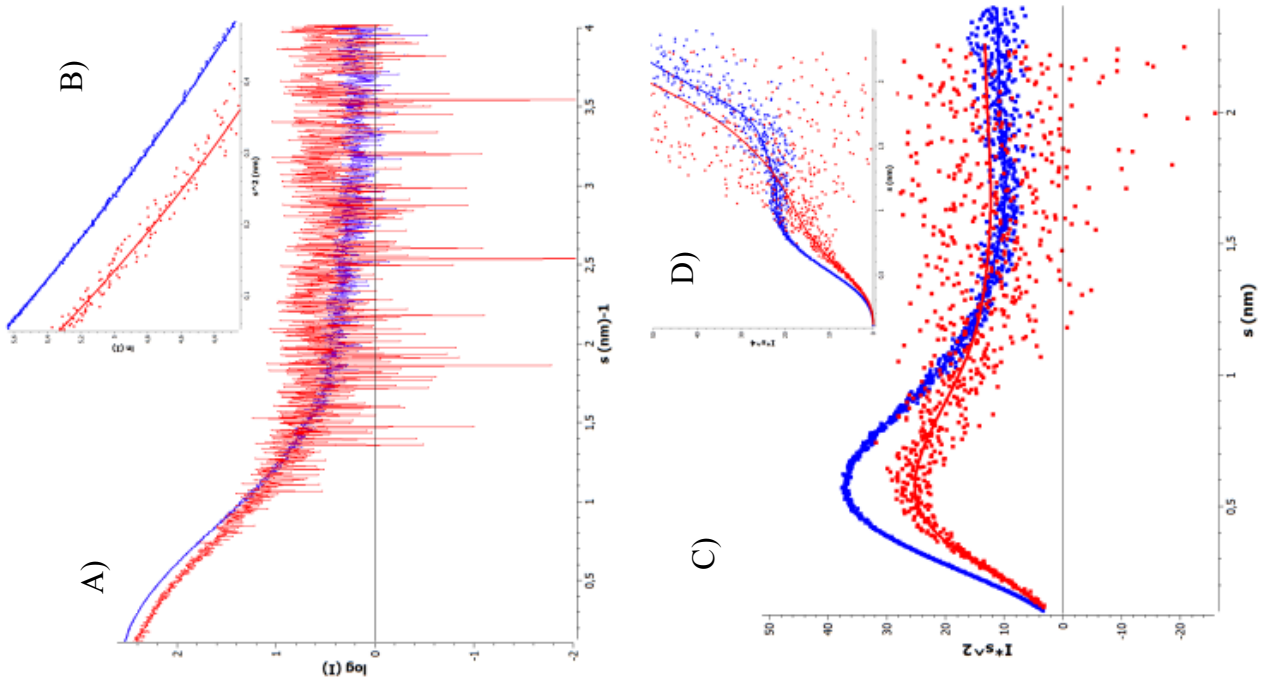


Fig.16 TmΔ88-502 construct measured in PETRA III. Two samples were measured. Both sample buffer solutions, were 10mM HEPES pH 7.5, 100mM NaCl, 100mM Glycine, 20µM FeSO4 and 2mM $CaCl_2$. The concentration of the HPLC-sample (blue) was 3.6 mg/ml and the batch sample was 0.5 mg/ml (red). The datasets were processed separately, scattering curve (A), Guinier plot (B), Kratky plot (C) and porod plot (D).

Experimentally the results could be improved increasing the number of measurements and with broader concentration series. The HPLC coupled measurement contains more data points improving the Guinier region linearity. The size exclusion in real time excludes the possibility of interfering aggregates. Therefore, the data collection with the PETRA III beamline was more successful experimentally. The data parameters seem to indicate that there is no difference whether the metal ions affect the enzyme stability. However, since the scattering curve indicates aggregation, the metals may cause the fold the become more flexible thus inducing aggregation.

The distance distribution functions P(r), show some right-handed shift and the distribution is non-symmetrical. The shift is seen as the concentration increases. The function suggests that

when the concentration increases in either condition, the protein becomes more aggregated. The HPCL-mode distribution function appears more symmetrical and appears as a compact monomer, and the quality of the data is much better than the batch samples.

In conclusion, according to these two datasets, the TmΔ88-502 construct is a monomer with a molecular weight of ~55kDa and globular with tail extension. The role of the metal ions according to the scattering profile and the distribution function is unclear. However, the interparticle artifacts can be observed, such as the increase of aggregation, causing instability. The increasing concentration seemed also to cause aggregation.

5.1.1 Molecular modeling

Based on the distribution function and the scattering curve, the shape and dimensions of the protein molecule were modeled using *Ab initio* simulated annealing approach. Two ATSAS software was used to model the molecules, DAMMIN (D. I. Svergun. 1999) and GASBOR (Petoukhov *et al.* 2012). The annealing simulation calculates different solutions comparing the real-space to the raw scattering curve trying to find the best fit in consequent iterations. The same dataset was run ten times in total, and the models and averaged using SUBCOMB (Kozin & Svergun. 2001) (**Fig.17**) The first five cycles were set up with minimal constraints and structure arguments. Additional five cycles were done with much slower and intense annealing with more arguments. For example, the predicted shape was fixed as a prolate, and a scaling factor was introduced into the iteration. The *Ab initio* models show that that the protein is an elongated monomer, where the N-terminus is extended.

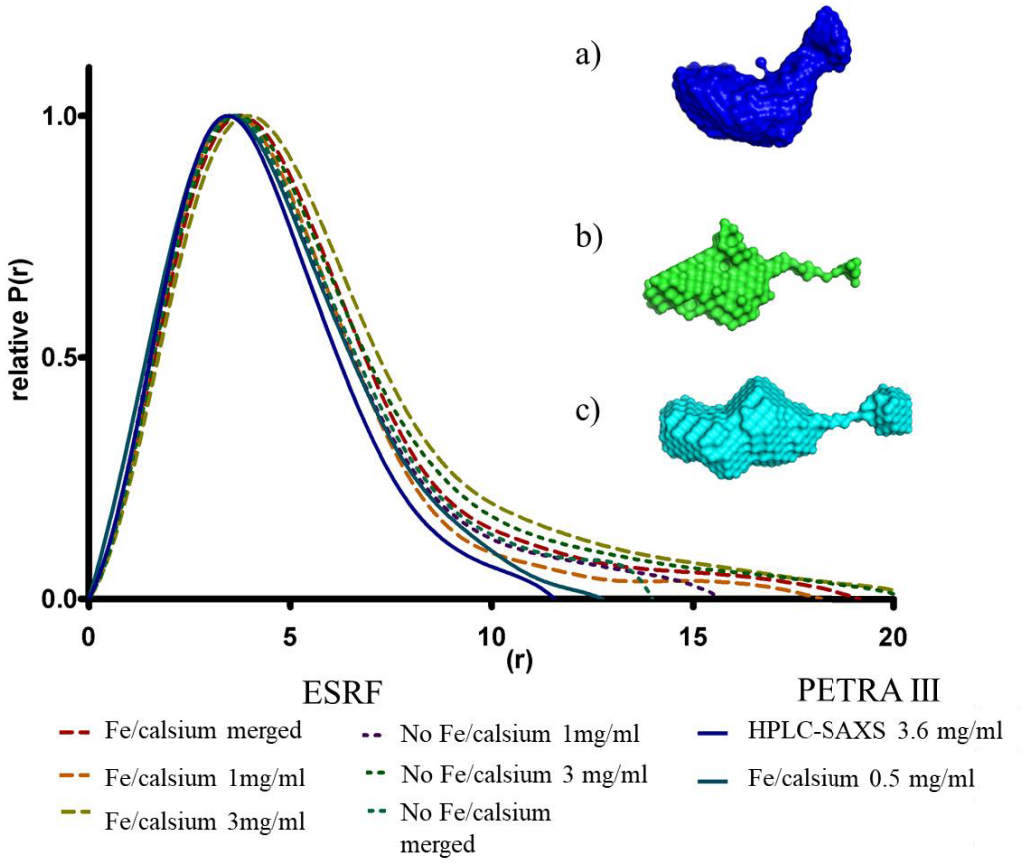


Fig.17 Distance distribution functions and dummy atom models of the TmΔ88-502 construct. The P(r) function displays the metal-/non-metal samples in from 15nm to 20 nm. The HPLC-mode sample and the 0.5mg/ml sample 11-12 nm. The magnitude of the area under the curve is different because the datasets were measured at two different beamlines. The P(r) functions were normalized. The dummy atom models are presented from the top in the following order, HPLC-sample (a), no metals (b), metals (c). The dummy atom models where constructed with gasbor. The final chi square values were 1.126, 0.987, 0.912 respectively.

Since the Cr-P4H has a 40% sequence similarity and many conserved regions, rigid body modeling was done. The crystal structure PDB:3GZE (Koski *et al.* 2009b) was fitted into the scattering curve using CRYSOL (D. Svergun *et al.* 1995), where the χ^2 score fit was 2.1. Since the Cr-P4H lacks the EF-hand part, the same calculation was done to the *Chlamydomonas* caltractin (PDB entry 3QRX) (Sosa Ldel *et al.* 2011). Additionally, HIF-P4Hs was also fitted

to the scattering, but the χ^2 was $1x10^2$-region. The amplitudes are used then used to fit the rigid body into the dummy atom model, and to build some missing regions with dummy atom chain based on the combined input sequence. The rigid body modeling was done using BUNCH (Petoukhov & Svergun. 2005). Since the construct is likely a monomer, instead of the Cr-P4H tetrameric complex structure a modified monomer structure was used 2JIG. The rigid body modeling was done similarly as the *Ab initio* modeling combining several annealing cycles. Distance arguments between the 2-OG domain of the Cr-P4H structure and the caltractin structure were 5Å. The gap between the β4β5 loop which is replaced by the EF-hand was measured in COOT (Emsley *et al.* 2010, Emsley & Cowtan. 2004). The distance argument was set in the BUNCH settings. Due to the elongation of the P4HTM model and the smaller size of Cr-P4H, the rigid body modeling proved to be too difficult. The missing regions in the Cr-P4H structure were too long to be modeled with dummy atoms. This causes the rigid body to overextend, and the model body does not fit in shape very well. The initial chi-square fit for the model was 2.1, but when the final annealing was completed, the value increased to 7, which is not any more meaningful.

5.2 Shorter constructs

Since the TmΔ88-502 construct suffered from insolubility issues, shorter constrtucts were generated. Two separate baculoviral systems were used to generate the constructs, the general baculovirus transfection system, and the Bac-to-Bac system (see materials and methods). Initially, the first attempt to clone the constructs failed, due to transfecting issues using the pVL1392 vector. The Bac-to-Bac system yielded expressing viruses, and the construct sequence was correct. Based on the secondary structure prediction, the constructs decided to start from positions E115, D130, G149, and Q187 hence referred to as TmΔ115-502. TmΔ130-502, TmΔ149-502, and TmΔ187-502 respectively. Some consideration was also used to determine the starting residue, based on the thermodynamics of short peptides (Bokhove *et al.* 2016, Varshavsky. 2011). The TmΔ115-502 construct starts at the first confidently predicted β-sheet. The following construct TmΔ130-502 starts from the helix before the EF-/EF-hand like domain. The last constructs start just before the EF-hand. The absence of the DSBH-fold supporting helix might cause the protein to unfold. Attempts were made to purify these constructs, with minimal yields. The elution profiles seemed corresponding to correct molecular

weight and first indicated that the TmΔ130-502 construct could appear as a dimer since the calculated MW of the construct is 45 kDa. The protein was visible on SDS-PAGE (**Fig**.**18**).

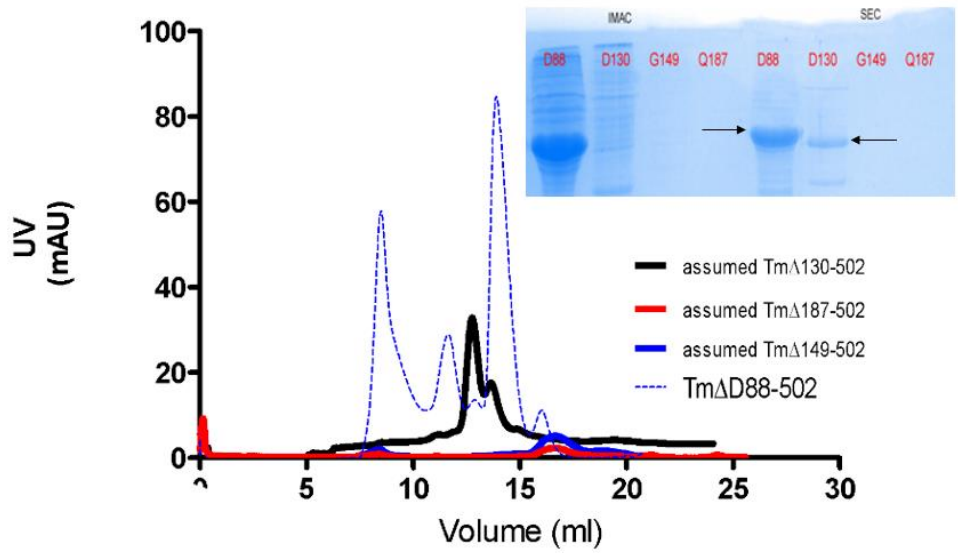


Fig.18 The elution profiles of purified shorter constructs. The retention volume of the TmΔ130-502 truncated variants appears to elute as a dimer and is observable on SDS-PAGE. MS analyses revealed that the sample belongs to an insect protein.

The uncoupled activity test was done, suggesting that there might be some activity, but most likely an artifact of the measurement. MALDI-TOF identification revealed that the sample belongs to an insect protein. The TmΔ88-502 was cloned successfully using the Bac-to-Bac system. However, in general, the construct generation failed due to technical issues with limited data, and time-related issues. Ideally, the shorter constructs were expected to limit the flexibility in the N-terminus and therefore increase the solubility. However, it is uncertain whether the absence of the supporting helix might have disrupted the fold. The shortest constructs TmΔ149-502 and TmΔ187-502 were not visible on the SDS-PAGE gel. Since the TmΔ88-502 was successfully generated, it is uncertain whether the cloning failed or that the shorter variants fall into the insoluble fraction, as a result of unfolding.

5.1 Effect of deglycosylation

Glycosylation as a post-translational modification usually affects the dynamics of the protein molecule, increasing the stability (Lee *et al.* 2015). For crystallization purposes, the glycans are often considered as an interfering factor, because it retains the protein in the soluble part of the

saturation curve (Chang *et al.* 2007). All tough, in glycosylation, does not necessary that the protein cant crystallizes, and in many cases, proteins can be crystallized with uncut glycans (Mesters & Hilgenfeld. 2007). The glycans can be cut using specific de-glycosylation enzymes such as endoglycosidase (Collin & Olsen. 2001, Maley *et al.* 1989, Mizuochi *et al.* 1984, Trastoy *et al.* 2018), often used with solubilizing enhancer such as maltose binding protein (Bokhove *et al.* 2016). The sample was deglycosylated after size exclusion chromatography. The mass difference of the de-glycosylation can be observed on SDS-PAGE (**Fig 19.**). The Solubility of the protein decreased but did not affect the protein folding properties or any significant changes to the secondary structure composition. The structural effects were determined to perform CD measurement for the de-glycosylated enzyme and comparing to the glycosylated enzyme (**Fig.20**). The 2-OG decarboxylation activity was not affected by the de-glycosylation, as the count rates were in the same level with the glycosylated protein. The activity supports the finding from the CD measurement that the 2-OG dioxygenase fold is not affected and it is still functional (Martina *et al.* 1998, Rasmussen. 1992).

The exact mas could not be determined since MALDI-TOF spectroscopy was not selective enough to determine the exact mass or the glycosylated residues.

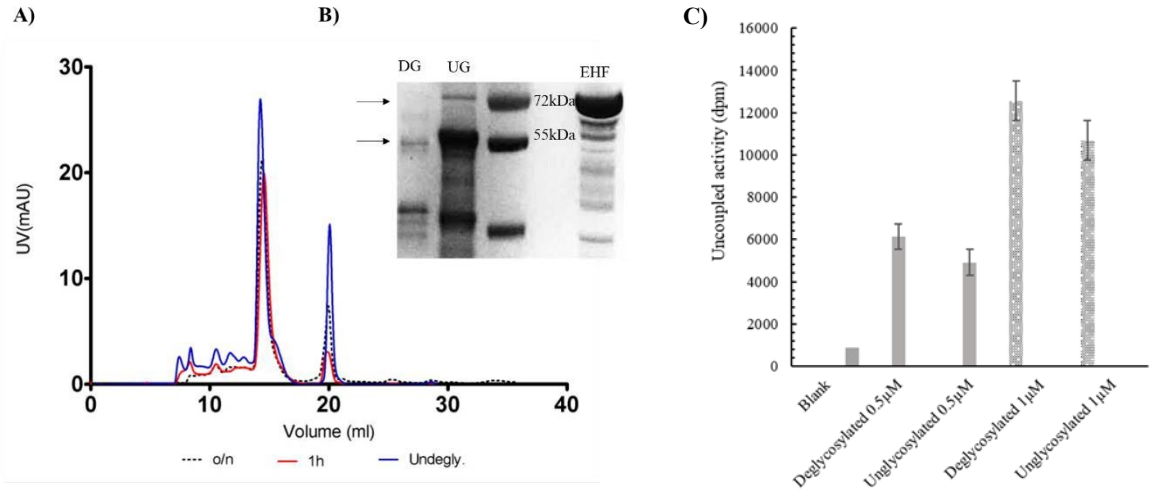


Fig.19 Purification of the deglycosylated TmΔ88-502 construct, which causes sample loss due to insolubility (A). The SDS-PAGE gel shows that the mass of the glucosylated protein (ug) is decreased to approximately 5kDa (dg) when endo hf is added (ehf) (B). The Endo HF fusion protein can be removed successfully with D-Maltose elution. The glycosylated sample has the fusion protein at 70 kDa level, but not in the deglycosylated sample. The activity of the undeglycosylated sample was measured (C). The deglycosylation does not affect the activity.

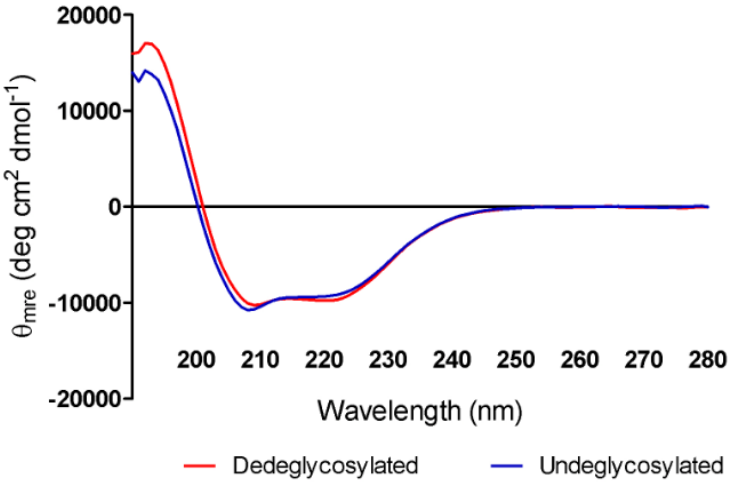


Fig.20 CD spectra- and activity test comparison against the deglycosylated protein. The CD curve shows that there is no change in the secondary structure after deglycosylation. The uncoupled activity shows that the enzyme is functional after deglycosylation.

Since the insolubility is a factor, the amount of soluble protein decreased, even more, when the purified sample was de-glycosylated was performed and repurified. Eventually, the quantities after de-glycosylation, gel-filtration, and sample concentration were low, causing the crystallization experiments to be limited. As a compromise, the de-glycosylating enzyme was added to the crystallization screen with a trace amount of endoglycosidase H, MBP fusion protein. The assuming that some fraction of the sample does not contain glycans and therefore pushing towards supersaturation more easily. Due to the difficulties of small yields and the insolubility issues, de-glycosylation was shown to be the ineffective approach for crystallization (**Fig.17**) Considering that the protein does not lose the structure or activity de-glycosylation. Due to the insolubility and low yields, the de-glycosylation could be used in crystallography when the yields are higher. The high insolubility could push the protein to crystallize easier.

5.2 The role of the calcium binding motif

The EF-hand is the most common calcium-binding motif which is part of hundreds of proteins. The first described EF-hand structure parvalbumin (Kretsinger & Nockolds. 1973)). The structure of EF-hand consists of the helix-loop-helix loop in the N-terminus order, helix-E, and helix-F. Hence the name EF-hand. The length of a single EF-hand is 30 residues long. The binding of calcium induces a conformational change, which causes motion in the motif. Since there are many different types of EF-hands with different affinities towards calcium, the magnitude of the motion depends on the class of the EF-hand. The role of EF-hand is involved in calcium buffering in the cytosol, and since calcium is a signal transduction molecule, EF-hand is involved in many signal transduction pathways (Wang & Schwarz. 2009).

In many cases, the function of EF-hand in the protein is to expose hydrophobic sites. Usually, catalytic enzyme residues reside in such hydrophobic sites. The most obvious example of EF-hand mediated substrate binding is seen on calmodulins (**Fig.21)** (Shifman *et al.* 2006) Calmodulins is ubiquitous signal transduction proteins which contain four EF-hands. Once calcium is bound to each hand, calmodulin is in open conformation enabling high affinity to the substrate protein such as kinases. The advantageous feature of EF-hand is that the helix groove is non-polar and flexible. The non-polar groove is not restricted to specific motifs in interacting proteins. The flexibility is caused by the loop between the helices, which allows the induce different motions depending on the substrate (Halling *et al.* 2016).

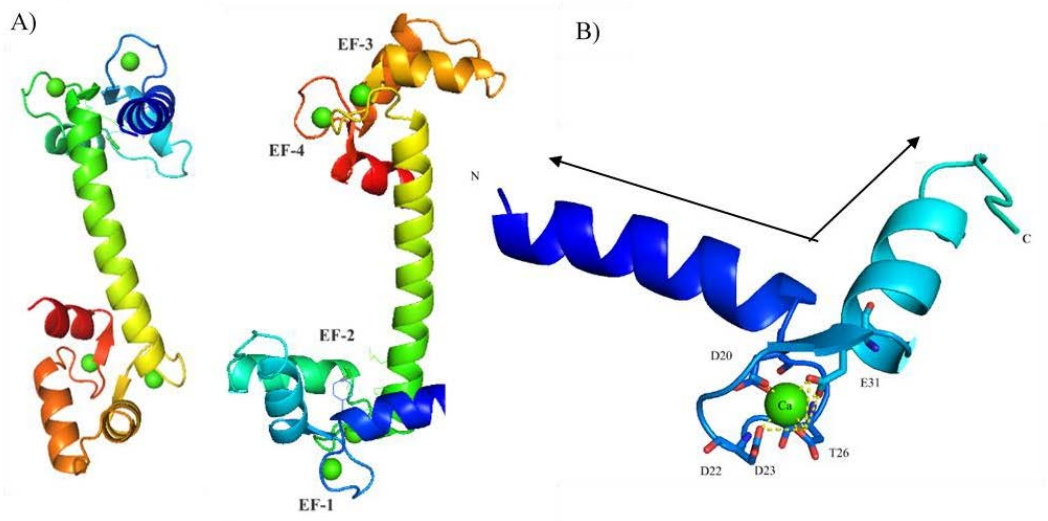


Fig.21 Crystal structure of human Calmodulin PDB entry 1CLL (Chattopadhyaya *et al.* 1992). Calmodulin contains four EF-hands with a connecting helix (A). A single EF-hand contains two helices similar as to thumb and the index finger (B). The arrows represent the motion included with calcium binding. The calcium atom is bound to one of the common binding motif DDDTE.

Since EF-hand is a very dynamic motif, which forms protein-protein interactions enabling conformational changes with their binding partners. The EF-hand contributes to the biological activity of the target protein (Meador *et al.* 1992, Zhang *et al.* 1995). The function of the EF-hand in P4HTM could relate to the substrate interaction. Both C-P4H and HIF-P4H substrates are transient substrates. Pro-collagen chains are disordered prior to the hydroxylation and HIFα belongs to the natively unfolded proteins. Since the sequence identity is closer to C-P4Hs than HIF-P4Hs, the P4HTM substrate is likely a secreted molecule. It should be noted that not all 2-OGDD substrates are proteins. The ER environment is crowded compared to other cellular compartments and contains partially folded proteins requiring post-translational modification inducing unfolded protein response (UPR) (Bravo *et al.* 2013, Velez Perez *et al.* 2013). P4HTM is not a ubiquitous protein but may adopt similar substrate binding properties. Like PDI (Pirneskoski *et al.* 2004, Ruddock *et al.* 1998), the P4HTM potent substrate could bound in a transient partially disordered state, where the EF-hand motif seals the substrate. An interesting factor is the oligomerization state *in vivo*? Does the protein appear as a monomer *in vitro* because it is not bound to the membrane? The possible mechanism and topology of P4HTM is proposed in (**Fig.20**)

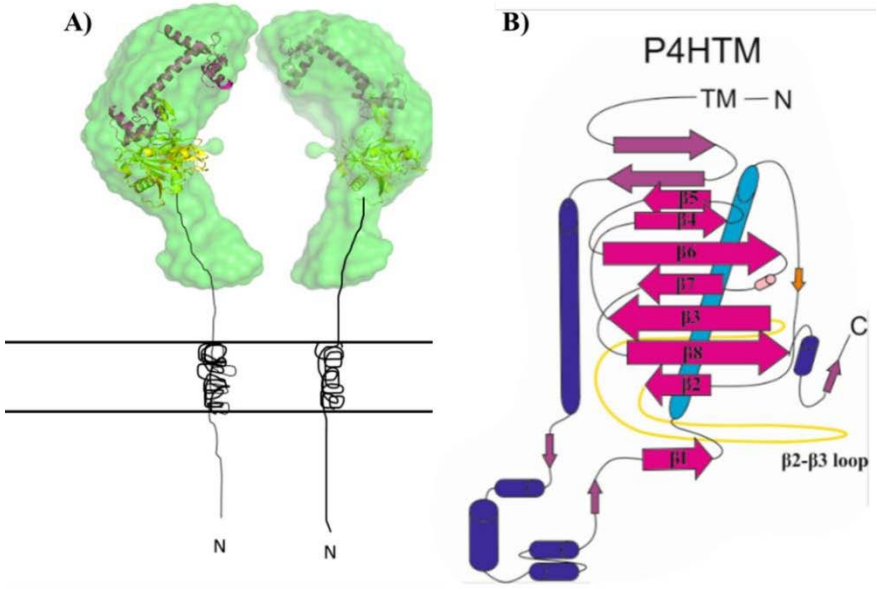


Fig.22 P4HTM appears as a dimer *in vivo*. The EF-hand motif interacts with the substrate which causes the protein to dimerize and bringing the substrate closer to the active site and once hydroxylated the EF-hands shifts back to open conformation (A). The model consists, SAXS data collected by the author and the calmodulin coordinates (1CLL) and from Cr-P4H (2JIG) (Koski et al. 2007) (Chattopadhyaya *et al.* 1992). The possible topology displays the β-loop replacement with EF-hand motif and the extended βII-βIII loop.

6. CONCLUSIONS

The study was done in this thesis characterizes P4HTM using spectroscopical techniques CD, DLS, and SEC-MALS. The kinetics of the 2-OG decarboxylation reaction was determined, to find Km values for oxygen and ferrous iron using radionuclide decay assays. Novel constructs were cloned into baculoviruses and attempted to purify. The structural properties were determined with small-angle X-ray scattering. Bioinformatics was used to perform molecular prediction models and compare gene homology.

The first low-resolution molecular solution structure model is proposed in this thesis. The characterization of the TmΔ88-502 construct, together with the SAXS data proofs that the transmembrane truncated slicing variant is a monomer and is active *in vitro*. The challenges of working with P4HTM arises mainly from the solubility. The data presented here show that the co-factors, Fe^{2+}- and Ca^{2+}-cations affects the stability of the enzyme. The solubility is highly dependent of the N-glycan asparagine residues. However, the activity or the stability of the structure is not affected. The solubility of the shorter consructs remained unclear and whether some of the constructs could be expressed. The uncoupled decarboxylation activity suggests that the enzyme is functional as a monomer. However, monomeric enzyme might not be a biologically active unit of the enzyme. The SAXS data from P4HTM fits better to collagen P4H crystal structure coordinates than HIF-P4H. P4HTM have been reported to hydroxylate HIF1α peptides, but not collagen peptides (Koivunen *et al.* 2007a) The *ab initio* model based on the SAXS data suggests that the active unit is globular with flexible extension towards the N-terminus. The extension suggests that full-length protein contains a flexible region between the transmembrane helix and the 2-OG dioxygenase domain. The flexibility could enable the dimerization *in vivo*. The EF-hands and the extended DSBD-loop might act as a tunnel trapping the substrate. Since the enzyme contains two EF-hand motifs, the enzyme could appear as a homodimer *in* vivo. The replacement of the nation-DSBH loop, e.g., β4-β5 in Cr-P4H, with the EF-hand domain, and the extension of the βII-βIII loop may suggest that the substrate is substantial or even disordered. The role of the EF-hand may be to close the conformation acting as a lid and sealing the active site (Hoeflich & Ikura. 2002). The result obtained from these experiments does not provide comprehensive details of the enzyme. The result shown here, act more as leads and suggestions, for future experimental approaches. Especially the binding affinity of the metal ions was not studied here.

7. REFERENCES

Agger K, Cloos PA, Christensen J, Pasini D, Rose S, Rappsilber J, Issaeva I, Canaani E, Salcini AE & Helin K. (2007) UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. Nature 449(7163): 731-734.

Aik W, McDonough MA, Thalhammer A, Chowdhury R & Schofield CJ. (2012) Role of the jelly-roll fold in substrate binding by 2-oxoglutarate oxygenases. Curr Opin Struct Biol 22(6): 691-700.

Anantharajan J, Koski MK, Kursula P, Hieta R, Bergmann U, Myllyharju J & Wierenga RK. (2013) The structural motifs for substrate binding and dimerization of the α subunit of collagen prolyl 4-hydroxylase. Structure 21(12): 2107.

Annunen P, Helaakoski T, Myllyharju J, Veijola J, Pihlajaniemi T & Kivirikko KI. (1997) Cloning of the human prolyl 4-hydroxylase alpha subunit isoform alpha(II) and characterization of the type II enzyme tetramer. The alpha(I) and alpha(II) subunits do not form a mixed alpha(I)alpha(II)beta2 tetramer. J Biol Chem 272(28): 17342-17348.

Beck I, Ramirez S, Weinmann R & Caro J. (1991) Enhancer element at the 3′-flanking region controls transcriptional response to hypoxia in the human erythropoietin gene. J Biol Chem 266(24): 5563–15566.

Bhatnagar RS, Prockop DJ & Rosenbloom J. (1967) Intracellular pool of unhydroxylated polypeptide precursors of collagen. Science 158(3800): 492-494.

Bieniossek C, Richmond TJ & Berger I. (2008) MultiBac: multigene baculovirus-based eukaryotic protein complex production. Curr Protoc Protein Sci Chapter 5: Unit 5.20.

Bishop T & Ratcliffe PJ. (2014) Signaling hypoxia by hypoxia-inducible factor protein hydroxylases: a historical overview and future perspectives. Hypoxia (Auckl). 2: 197-213.

Blaho D. (2008) X-Ray Crystallography of Biomacromolecules. Yale J Biol Med 81(1): 51-52.

Blasiak LC & Drennan CL. (2009) Structural perspective on enzymatic halogenation. Acc Chem Res 42(1): 147-155.

Bokhove M, Hosseini SA, Saito T, Dioguardi E, Gegenschatz-Schmid K, Nishimura K, Raj I, de Sanctis D, Han L & Jovine L. (2016) Easy mammalian expression and crystallography of maltose-binding protein-fused human proteins. J Struct Biol 194(1): 1-7.

Borowski T, Bassan A & Siegbahn PEM. (2004) Mechanism of Dioxygen Activation in 2-Oxoglutarate-Dependent Enzymes: A Hybrid DFT Study. Chemistry – A European Journal 10(4): 1031-1041.

Bravo R, Parra V, Gatica D, Rodriguez AE, Torrealba N, Paredes F, Wang ZV, Zorzano A, Hill JA, Jaimovich E, Quest AFG & Lavandero S. (2013) Endoplasmic reticulum and the unfolded protein response: dynamics and metabolic integration. International review of cell and molecular biology 301: 215-290.

Brodsky B & Persikov AV. (2005) The molecular structure of the collagen triple helix. Adv Protein Chem 70: 301-339.

Bruick RK & McKnight SL. (2001) A Conserved Family of Prolyl-4-Hydroxylases That Modify HIF. Science 294(5545): 1337.

Chang VT, Crispin M, Aricescu AR, Harvey DJ, Nettleship JE, Fennelly JA, Yu C, Boles KS, Evans EJ, Stuart DI, Dwek RA, Jones EY, Owens RJ & Davis SJ. (2007) Glycoprotein structural genomics: solving the glycosylation problem. Structure (London, England : 1993) 15(3): 267-273.

Chattopadhyaya R, Meador WE, Means AR & Quiocho FA. (1992) Calmodulin structure refined at 1.7 A resolution. J Mol Biol 228(4): 1177-1192.

Chen Z, Zang J, Kappler J, Hong X, Crawford F, Wang Q, Lan F, Jiang C, Whetstine J, Dai S, Hansen K, Shi Y & Zhang G. (2007) Structural basis of the recognition of a methylated histone tail by JMJD2A. Proc Natl Acad Sci U S A 104(26): 10818-10823.

Chowdhury R, McDonough MA, Mecinovic J, Loenarz C, Flashman E, Hewitson KS, Domene C & Schofield CJ. (2009) Structural basis for binding of a hypoxia-inducible factor to the oxygen-sensing prolyl hydroxylases. Structure 17(7): 981-989.

Chowdhury R, Sekirnik R, Brissett NC, Krojer T, Ho CH, Ng SS, Clifton IJ, Ge W, Kershaw NJ, Fox GC, Muniz JRC, Vollmar M, Phillips C, Pilka ES, Kavanagh KL, von Delft F, Oppermann U, McDonough MA, Doherty AJ & Schofield CJ. (2014) Ribosomal oxygenases are structurally conserved from prokaryotes to humans. Nature 510(7505): 422-426.

Clifton IJ, McDonough MA, Ehrismann D, Kershaw NJ, Granatino N & Schofield CJ. (2006) Structural studies on 2-oxoglutarate oxygenases and related double-stranded beta-helix fold proteins. J Inorg Biochem 100(4): 644-669.

Coleman ML, McDonough MA, Hewitson KS, Coles C, Mecinovic J, Edelmann M, Cook KM, Cockman ME, Lancaster DE, Kessler BM, Oldham NJ, Ratcliffe PJ & Schofield CJ. (2007) Asparaginyl hydroxylation of the Notch ankyrin repeat domain by factor inhibiting hypoxia-inducible factor. J Biol Chem 282(33): 24027-24038.

Collin M & Olsen A. (2001) EndoS, a novel secreted protein from Streptococcus pyogenes with endoglycosidase activity on human IgG. EMBO J 20(12): 3046-3055.

Costas M, Mehn MP, Jensen MP & Que L. (2004) Dioxygen Activation at Mononuclear Nonheme Iron Active Sites: Enzymes, Models, and Intermediates. Chem Rev 104(2): 939-986.

Couture JF, Collazo E, Ortiz-Tello PA, Brunzelle JS & Trievel RC. (2007) Specificity and mechanism of JMJD2A, a trimethyllysine-specific histone demethylase. Nat Struct Mol Biol 14(8): 689-695.

Croes K, Van Veldhoven PP, Mannaerts GP & Casteels M. (1997) Production of formyl-CoA during peroxisomal alpha-oxidation of 3-methyl-branched fatty acids. FEBS Lett 407(2): 197-200.

Cserzo M, Wallin E, Simon I, von Heijne G & Elofsson A. (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng 10(6): 673-676.

Culver C, Sundqvist A, Mudie S, Melvin A, Xirodimas D & Rocha S. (2010) Mechanism of hypoxia-induced NF-kappaB. Mol Cell Biol 30(20): 4901-4921.

Dann CE,3rd, Bruick RK & Deisenhofer J. (2002) Structure of factor-inhibiting hypoxia-inducible factor 1: An asparaginyl hydroxylase involved in the hypoxic response pathway. Proc Natl Acad Sci U S A 99(24): 15351-15356.

Dao JH, Kurzeja RJM, Morachis JM, Veith H, Lewis J, Yu V, Tegley CM & Tagari P. (2009) Kinetic characterization and identification of a novel inhibitor of hypoxia-inducible factor prolyl hydroxylase 2 using a time-resolved fluorescence resonance energy transfer-based assay technology. Anal Biochem 384(2): 213-223.

Ehrismann D, Flashman E, Genn D, Mathioudakis N, Hewitson K, Ratcliffe P & Schofield C. (2007) Studies on the activity of the hypoxia-inducible-factor hydroxylases using an oxygen consumption assay. Biochem J 401(1): 227.

Ellgaard L & Ruddock LW. (2005) The human protein disulphide isomerase family: substrate interactions and functional properties. EMBO Rep 6(1): 28-32.

Emsley P, Lohkamp B, Scott WG & Cowtan K. (2010) Features and development of Coot. Acta Crystallographica Section D 66(4): 486-501.

Emsley P & Cowtan K. (2004) Coot: model-building tools for molecular graphics. Acta Crystallographica Section D 60(12-1): 2126-2132.

Epstein AC, Gleadle JM, McNeill LA, Hewitson KS, O'Rourke J, Mole DR, Mukherji M, Metzen E, Wilson MI, Dhanda A, Tian YM, Masson N, Hamilton DL, Jaakkola P, Barstead R, Hodgkin J, Maxwell PH, Pugh CW, Schofield CJ & Ratcliffe PJ. (2001) C. elegans EGL-9 and mammalian homologs define a family of dioxygenases that regulate HIF by prolyl hydroxylation. Cell. 107(1): 43-54.

Fersht A. (1999) Structure and Mechanism in Protein Science. New York: Freeman.

Fischer E. (1902) Über eine neue Aminosäure aus Leim. Chem Ber. 35: 2660-2665.

Flashman E, Hoffart LM, Hamed RB, Bollinger JM, Krebs C & Schofield CJ. (2010) Evidence for the slow reaction of hypoxia-inducible factor prolyl hydroxylase 2 with oxygen. The FEBS journal 277(19): 4089-4099.

Goldberg B & Green H. (1968) The synthesis of collagen and protocollagen hydroxylase by fibroblastic and nonfibroblastic cell lines. Proc Natl Acad Sci U S A 59(4): 1110-1115.

Good NE, Winget GD, Winter W, Connolly TN, Izawa S & Singh RMM. (1966) Hydrogen Ion Buffers for Biological Research*. Biochemistry (N Y) 5(2): 467-477.

Gordan JD & Simon MC. (2007) Hypoxia-inducible factors: central regulators of the tumor phenotype. Curr Opin Genet Dev 17(1): 71-77.

Gordon MK & Hahn RA. (2010) Collagens. Cell Tissue Res 339(1): 247-257.

Gorres KL & Raines RT. (2010) Prolyl 4-hydroxylase. Crit Rev Biochem Mol Biol 45(2): 106-124.

Gräslund S, Nordlund P, Weigelt J, Hallberg BM, Bray J, Gileadi O, Knapp S, Oppermann U, Arrowsmith C, Hui R, Ming J, dhe-Paganon S, Park H, Savchenko A, Yee A, Edwards A, Vincentelli R, Cambillau C, Kim R, Kim S, Rao Z, Shi Y, Terwilliger TC, Kim C, Hung L, Waldo GS, Peleg Y, Albeck S, Unger T, Dym O, Prilusky J, Sussman JL, Stevens RC, Lesley SA, Wilson IA, Joachimiak A, Collart F, Dementieva I, Donnelly MI, Eschenfeldt WH, Kim Y, Stols L, Wu R, Zhou M, Burley SK, Emtage JS, Sauder JM, Thompson D, Bain K, Luz J, Gheyi T, Zhang F, Atwell S, Almo SC, Bonanno JB, Fiser A, Swaminathan S, Studier FW, Chance MR, Sali A, Acton TB, Xiao R, Zhao L, Ma LC, Hunt JF, Tong L, Cunningham K, Inouye M, Anderson S, Janjua H, Shastry R, Ho CK, Wang D, Wang H, Jiang M, Montelione GT, Stuart DI, Owens RJ, Daenke S, Schütz A, Heinemann U, Yokoyama S & Büssow K. (2008) Protein production and purification. Nature methods 5(2): 135-146.

Greenfield NJ. (2006) Using circular dichroism spectra to estimate protein secondary structure. Nature protocols 1(6): 2876-2890.

Grzyska PK, Appelman EH, Hausinger RP & Proshlyakov DA. (2010) Insight into the mechanism of an iron dioxygenase by resolution of steps following the FeIV=HO species. Proc Natl Acad Sci U S A 107(9): 3982-3987.

Guarente L. (2011) Sirtuins, aging, and metabolism. Cold Spring Harb Symp Quant Biol 76: 81-90.

Halling DB, Liebeskind BJ, Hall AW & Aldrich RW. (2016) Conserved properties of individual Ca2+-binding sites in calmodulin. Proc Natl Acad Sci U S A 113(9): E1225.

Hanauske-Abel H & Günzler V. (1982) A stereochemical concept for the catalytic mechanism of prolylhydroxylase: Applicability to classification and design of inhibitors. J Theor Biol 94(2): 421-455.

Hausinger RP. (2004) FeII/alpha-ketoglutarate-dependent hydroxylases and related enzymes. Crit Rev Biochem Mol Biol 39(1): 21-68.

Hendrickson WA & Ogata CM. (1997) 28] Phase determination from multiwavelength anomalous diffraction measurements. Methods Enzymol 276: 494-523.

Hernández-Ortega A, Quesne MG, Bui S, Heyes DJ, Steiner RA, Scrutton NS & de Visser SP. (2015) Catalytic Mechanism of Cofactor-Free Dioxygenases and How They Circumvent Spin-Forbidden Oxygenation of Their Substrates. J Am Chem Soc 137(23): 7474-7487.

Hewitson KS, Granatino N, Welford RW, McDonough MA & Schofield CJ. (2005) Oxidation by 2-oxoglutarate oxygenases: non-haem iron systems in catalysis and signalling. Philos Trans A Math Phys Eng Sci 363(1829): 40.

Hieta R, Kukkola L, Permi P, Pirila P, Kivirikko KI, Kilpelainen I & Myllyharju J. (2003) The peptide-substrate-binding domain of human collagen prolyl 4-hydroxylases. Backbone assignments, secondary structure, and binding of proline-rich peptides. J Biol Chem 278(37): 34966-34974.

Hieta R & Myllyharju J. (2002) Cloning and characterization of a low molecular weight prolyl 4-hydroxylase from Arabidopsis thaliana. Effective hydroxylation of proline-rich, collagen-like, and hypoxia-inducible transcription factor alpha-like peptides. J Biol Chem 277(26): 23965-23971.

Hirsila M, Koivunen P, Gunzler V, Kivirikko KI & Myllyharju J. (2003) Characterization of the human prolyl 4-hydroxylases that modify the hypoxia-inducible factor. J Biol Chem 278(33): 30772-30780.

Hoeflich KP & Ikura M. (2002) Calmodulin in Action: Diversity in Target Recognition and Activation Mechanisms. Cell 108(6): 739-742.

Hoffart LM, Barr EW, Guyer RB, Bollinger JM,Jr & Krebs C. (2006) Direct spectroscopic detection of a C-H-cleaving high-spin Fe(IV) complex in a prolyl-4-hydroxylase. Proc Natl Acad Sci U S A 103(40): 14738-14743.

Hu CJ, Wang LY, Chodosh LA, Keith B & Simon MC. (2003) Differential roles of hypoxia-inducible factor 1alpha (HIF-1alpha) and HIF-2alpha in hypoxic gene regulation. Mol Cell Biol 23(24): 9361-9374.

Hu L, Li Z, Cheng J, Rao Q, Gong W, Liu M, Shi YG, Zhu J, Wang P & Xu Y. (2013) Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation. Cell 155(7): 1545-1555.

Huang J, Zhao Q, Mooney SM & Lee FS. (2002) Sequence determinants in hypoxia-inducible factor-1alpha for hydroxylation by the prolyl hydroxylases PHD1, PHD2, and PHD3 . J Biol Chem. 277(42): 39792-800.

Huang LE, Gu J, Schau M & Bunn HF. (1998) Regulation of hypoxia-inducible factor 1alpha is mediated by an O2-dependent degradation domain via the ubiquitin-proteasome pathway. Proc Natl Acad Sci U S A. 95(14): 7987-92.

Hudson DM & Eyre DR. (2013) Collagen prolyl 3-hydroxylation: a major role for a minor post-translational modification? Connect Tissue Res 54(4-5): 245-251.

Huggins ML. (1945) Theory of X-ray Diffraction in Crystals (Zachariasen, William H.). J Chem Educ 22(7): 364.

Hutton JJ,Jr, Trappel AL & Udenfriend S. (1966) Requirements for alpha-ketoglutarate, ferrous ion and ascorbate by collagen proline hydroxylase. Biochem Biophys Res Commun 24(2): 179-184.

Hutton JJ,Jr & Udenfriend S. (1966) Soluble collagen proline hydroxylase and its substrates in several animal tissues. Proc Natl Acad Sci U S A 56(1): 198-202.

Hyvärinen J, Parikka M, Sormunen R, Rämet M, Tryggvason K, Kivirikko KI, Myllyharju J & Koivunen P. (2010) Deficiency of a transmembrane prolyl 4-hydroxylase in the zebrafish leads to basement membrane defects and compromised kidney function. The Journal of biological chemistry 285(53): 42023-42032.

Islam MS, Leissing TM, Chowdhury R, Hopkinson RJ & Schofield CJ. (2018) 2-Oxoglutarate-Dependent Oxygenases. Annu Rev Biochem 87: 585-620.

Ivan M, Haberberger T, Gervasi DC, Michelson KS, Günzler V, Kondo K, Yang H, Sorokina I, Conaway RC, Conaway JW & Kaelin WG. (2002) Biochemical purification and pharmacological inhibition of a mammalian prolyl hydroxylase acting on hypoxia-inducible factor. Proc Natl Acad Sci U S A. 99(21): 13459-64.

Ivan M, Kondo K, Yang H, Kim W, Valiando J, Ohh M, Salic A, Asara JM, Lane WS & Kaelin WG. (2001) HIFalpha targeted for VHL-mediated destruction by proline hydroxylation: implications for O2 sensing. Science 292(5516): 464-8.

Jaakkola P, Mole DR, Tian YM, Wilson MI, Gielbert J, Gaskell SJ, von Kriegsheim A, Hebestreit HF, Mukherji M, Schofield CJ, Maxwell PH, Pugh CW & Ratcliffe PJ. (2001) Targeting of HIF-alpha to the von Hippel-Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. Science 292(5516): 72.

Jiang BH, Rue EA, Wang GL, Roe R & Semenza GL. (1996) Dimerization, DNA binding, and transactivation properties of hypoxia-inducible factor 1. J Biol Chem 271: 17771-17778.

Kaasinen E, Rahikkala E, Koivunen P, Miettinen S, Wamelink MMC, Aavikko M, Palin K, Myllyharju J, Moilanen JS, Pajunen L, Karhu A & Aaltonen LA. (2014) Clinical characterization, genetic mapping and whole-genome sequence analysis of a novel autosomal recessive intellectual disability syndrome. European Journal of Medical Genetics 57(10): 543-551.

Kaelin WG & Ratcliffe PJ. (2008) Oxygen Sensing by Metazoans: The Central Role of the HIF Hydroxylase Pathway. Molecular Cell 30(4): 393-402.

Kant R, Bali A, Singh N & Jaggi AS. (2013) Prolyl 4 hydroxylase: a critical target in the pathophysiology of diseases. The Korean journal of physiology & pharmacology : official journal of the Korean Physiological Society and the Korean Society of Pharmacology 17(2): 111-120.

Karlin KD. (1993) Metalloenzymes, structural motifs, and inorganic models. Science 261(5122): 701.

Kato M, Araiso Y, Noma A, Nagao A, Suzuki T, Ishitani R & Nureki O. (2011) Crystal structure of a novel JmjC-domain-containing protein, TYW5, involved in tRNA modification. Nucleic Acids Res 39(4): 1576-1585.

Kelley LA, Mezulis S, Yates CM, Wass MN & Sternberg MJE. (2015) The Phyre2 web portal for protein modeling, prediction, and analysis. Nature Protocols 10(6): 845-858.

Kemmink J, Darby NJ, Dijkstra K, Nilges M & Creighton TE. (1996) Structure determination of the N-terminal thioredoxin-like domain of protein disulfide isomerase using multidimensional heteronuclear 13C/15N NMR spectroscopy. Biochemistry 35(24): 7684-7691.

Kivirikko KI, Myllyla R & Pihlajaniemi T. (1989) Protein hydroxylation: prolyl 4-hydroxylase, an enzyme with four cosubstrates and a multifunctional subunit. FASEB J 3(5): 1609-1617.

Kivirikko KI & Pihlajaniemi T. (1998) Collagen hydroxylases and the protein disulfide isomerase subunit of prolyl 4-hydroxylases. Adv Enzymol Relat Areas Mol Biol 72: 325-398.

Kivirikko KI & Prockop DJ. (1967) ENZYMATIC HYDROXYLATION OF PROLINE AND LYSINE IN PROTOCOLLAGEN. Proc Natl Acad Sci USA 57(3): 782-789.

Kivirikko KI & Prockop DJ. (1972) Partial purification and characterization of protocollagen lysine hydroxylase from chick embryos. BBA - Enzymology 258(2): 366-379.

Koivunen P, Hirsila M, Kivirikko KI & Myllyharju J. (2006) The length of peptide substrates has a marked effect on hydroxylation by the hypoxia-inducible factor prolyl 4-hydroxylases. J Biol Chem 281(39): 28712-28720.

Koivunen P, Salo KE, Myllyharju J & Ruddock LW. (2005) Three binding sites in protein-disulfide isomerase cooperate in collagen prolyl 4-hydroxylase tetramer assembly. J Biol Chem 280(7): 5227-5235.

Koivunen P, Tiainen P, Hyvärinen J, Williams KE, Sormunen R, Klaus SJ, Kivirikko KI & Myllyharju J. (2007a) An Endoplasmic Reticulum Transmembrane Prolyl 4-Hydroxylase Is Induced by Hypoxia and Acts on Hypoxia-inducible Factor α. J Biol Chem 282(42): 30544.

Koivunen P, Hirsilae M, Remes AM, Hassinen IE, Kivirikko KI & Myllyharju J. (2007b) Inhibition of Hypoxia-inducible Factor (HIF) Hydroxylases by Citric Acid Cycle Intermediates: Possible links between cell metabolism and stabilization of HIF. Journal of Biological Chemistry 282(7): 4524-4532.

Koski MK, Hieta R, Hirsilä M, Rönkä A, Myllyharju J & Wierenga RK. (2009a) The crystal structure of an algal prolyl 4-hydroxylase complexed with a proline-rich peptide reveals a novel buried tripeptide binding motif. J Biol Chem. 2009 284(37): 25290-301.

Koski MK, Anantharajan J, Kursula P, Dhavala P, Murthy AV, Bergmann U, Myllyharju J & Wierenga RK. (2017) Assembly of the elongated collagen prolyl 4-hydroxylase alpha2beta2 heterotetramer around a central alpha2 dimer. Biochem J 474(5): 751-769.

Koski MK, Hieta R, Bollner C, Kivirikko KI, Myllyharju J & Wierenga RK. (2007) The active site of an algal prolyl 4-hydroxylase has large structural plasticity. J Biol Chem 282(51): 37112-37123.

Koski MK, Hieta R, Hirsila M, Ronka A, Myllyharju J & Wierenga RK. (2009b) The crystal structure of an algal prolyl 4-hydroxylase complexed with a proline-rich peptide reveals a novel buried tripeptide binding motif. J Biol Chem 284(37): 25290-25301.

Kozin MB & Svergun DI. (2001) Automated matching of high- and low-resolution structural models. Journal of Applied Crystallography 34(1): 33-41.

Krebs C, Galonic Fujimori D, Walsh CT & Bollinger JM,Jr. (2007) Non-heme Fe(IV)-oxo intermediates. Acc Chem Res 40(7): 484-492.

Kretsinger RH & Nockolds CE. (1973) Carp muscle calcium-binding protein. II. Structure determination and general description. J Biol Chem 248(9): 3313-3326.

Kukkola L, Koivunen P, Pakkanen O, Page AP & Myllyharju J. (2004) Collagen prolyl 4-hydroxylase tetramers and dimers show identical decreases in Km values for peptide substrates with increasing chain length: mutation of one of the two catalytic sites in the tetramer inactivates the enzyme by more than half. J Biol Chem 279(18): 18656-18661.

Kurowski MA, Bhagwat AS, Papaj G & Bujnicki JM. (2003) Phylogenomic identification of five new human homologs of the DNA repair enzyme AlkB. BMC Genomics 4(1): 48.

La Verde V, Dominici P & Astegno A. (2017) Determination of Hydrodynamic Radius of Proteins by Size Exclusion Chromatography. BIO-PROTOCOL 7(8).

Laitala A, Aro E, Walkinshaw G, Maki JM, Rossi M, Heikkila M, Savolainen ER, Arend M, Kivirikko KI, Koivunen P & Myllyharju J. (2012) Transmembrane prolyl 4-hydroxylase is a fourth prolyl 4-hydroxylase regulating EPO production and erythropoiesis. Blood 120(16): 3336-3344.

Lee HS, Qi Y & Im W. (2015) Effects of N-glycosylation on protein conformation and dynamics: Protein Data Bank analysis and molecular dynamics simulation study. Scientific reports 5: 8926.

Leinonen H, Rossi M, Salo A,M., Tiainen P, Hyvärinen J, Pitkänen M, Sormunen R, Miinalainen I, Zhang C, Soininen R, Kivirikko K,I., Koskelainen A, Tanila H, Myllyharju J & Koivunen P. (2016) Lack of P4H-TM in mice results in age-related retinal and renal alterations. Human Molecular Genetics 25(17): 3810-3823.

Lieb ME, Menzies K, Moschella MC, Ni R & Taubman MB. (2002) Mammalian EGLN genes have distinct patterns of mRNA expression and regulation. Biochem Cell Biol 80(4): 421-426.

Liu X, Wang L, Zhao K, Thompson PR, Hwang Y, Marmorstein R & Cole PA. (2008) The structural basis of protein acetylation by the p300/CBP transcriptional coactivator. Nature 451(7180): 846-850.

Loenarz C & Schofield CJ. (2011) Physiological and biochemical aspects of hydroxylations and demethylations catalyzed by human 2-oxoglutarate oxygenases. Trends in Biochemical Sciences 36(1): 7-18.

Longbotham JE, Levy C, Johannissen LO, Tarhonskaya H, Jiang S, Loenarz C, Flashman E, Hay S, Schofield CJ & Scrutton NS. (2015) Structure and Mechanism of a Viral Collagen Prolyl Hydroxylase. Biochemistry 54(39): 6093-6105.

Lundberg M, Siegbahn PEM & Morokuma K. (2008) The Mechanism for Isopenicillin N Synthase from Density-Functional Modeling Highlights the Similarities with Other Enzymes in the 2-His-1-carboxylate Family. Biochemistry 47(3): 1031-1042.

Majamaa K, Hanauske-Abel HM, Gunzler V & Kivirikko KI. (1984) The 2-oxoglutarate binding site of prolyl 4-hydroxylase. Identification of distinct subsites and evidence for 2-oxoglutarate decarboxylation in a ligand reaction at the enzyme-bound ferrous ion. Eur J Biochem 138(2): 239-245.

Malaby AW, Chakravarthy S, Irving TC, Kathuria SV, Bilsel O & Lambright DG. (2015) Methods for analysis of size-exclusion chromatography-small-angle X-ray scattering and reconstruction of protein scattering. J Appl Crystallogr 48(Pt 4): 1102-1113.

Maley F, Trimble RB, Tarentino AL & Plummer TH. (1989) Characterization of glycoproteins and their associated oligosaccharides through the use of endoglycosidases. Analytical Biochemistry 180(2): 195-204.

Marini JC, Cabral WA, Barnes AM & Chang W. (2007) Components of the collagen prolyl 3-hydroxylation complex are crucial for normal bone development. Cell Cycle 6(14): 1675-1681.

Martina JA, Daniotti JL & Maccioni HJ. (1998) Influence of N-glycosylation and N-glycan trimming on the activity and intracellular traffic of GD3 synthase. J Biol Chem 273(6): 3725-3731.

Martinez S & Hausinger RP. (2015) Catalytic Mechanisms of Fe(II)- and 2-Oxoglutarate-dependent Oxygenases. The Journal of biological chemistry 290(34): 20702-20711.

McCall KA, Huang C & Fierke CA. (2000) Function and mechanism of zinc metalloenzymes. J Nutr 130(5S Suppl): 46S.

McDonough MA, Loenarz C, Chowdhury R, Clifton IJ & Schofield CJ. (2010) Structural studies on human 2-oxoglutarate dependent oxygenases. Curr Opin Struct Biol 20(6): 659-672.

McDonough MA, Li V, Flashman E, Chowdhury R, Mohr C, Liénard BMR, Zondlo J, Oldham NJ, Clifton IJ, Lewis J, McNeill LA, Kurzeja RJM, Hewitson KS, Yang E, Jordan S, Syed RS & Schofield CJ. (2006) Cellular oxygen sensing: Crystal structure of

hypoxia-inducible factor prolyl hydroxylase (PHD2). Proc Natl Acad Sci USA 103(26): 9814-9819.

McPherson A & Gavira JA. (2013) Introduction to protein crystallization. Acta crystallographica.Section F, Structural biology communications 70: 2-20.

Meador WE, Means AR & Quiocho FA. (1992) Target enzyme recognition by calmodulin: 2.4 A structure of a calmodulin-peptide complex. Science 257(5074): 1251-1255.

Mehta N, Porterfield M, Struwe WB, Heiss C, Azadi P, Rudd PM, Tiemeyer M & Aoki K. (2016) Mass Spectrometric Quantification of N-Linked Glycans by Reference to Exogenous Standards. Journal of proteome research 15(9): 2969-2980.

Mesters JR & Hilgenfeld R. (2007) Protein Glycosylation, Sweet to Crystal Growth? Crystal Growth & Design 7(11): 2251-2253.

Michiels C. (2004) Physiological and Pathological Responses to Hypoxia. American Journal of Pathology 164(6): 1875-1882.

Min JH, Yang H, Ivan M, Gertler F, Kaelin WG,Jr & Pavletich NP. (2002) Structure of an HIF-1alpha -pVHL complex: hydroxyproline recognition in signaling. Science 296(5574): 1886-1889.

Mizuochi T, Amano J & Kobata A. (1984) New evidence of the substrate specificity of endo-beta-N-acetylglucosaminidase D. J Biochem 95(4): 1209-1213.

Moroz E, Carlin S, Dyomina K, Burke S, Thaler HT, Blasberg R & Serganova I. (2009) Real-time imaging of HIF-1alpha stabilization and degradation. PLoS One 4(4): e5077.

Mouw JK, Ou G & Weaver VM. (2014) Extracellular matrix assembly: a multiscale deconstruction. Nature Reviews Molecular Cell Biology 15: 771.

Murthy AV. (2018) Structure-Function Studies of Human Collagen Prolyl 4-Hydroxylase Isoform II and its Peptide-Substrate-Binding Domain. Univercity of Oulu.

Murthy AV, Sulu R, Koski MK, Tu H, Anantharajan J, Sah-Teli S, Myllyharju J & Wierenga RK. (2018a) Structural enzymology binding studies of the peptide-substrate-binding domain of human collagen prolyl 4-hydroxylase (type-II): High affinity peptides have a PxGP sequence motif. Protein Science 27(9): 1692-1703.

Murthy AV, Sulu R, Koski MK, Tu H, Anantharajan J, Sah-Teli S, Myllyharju J & Wierenga RK. (2018b) Structural enzymology binding studies of the peptide-substrate-binding domain of human collagen prolyl 4-hydroxylase (type-II): High affinity peptides have a PxGP sequence motif. Protein Science 27(9): 1692-1703.

Myllyharju J. (2003) Prolyl 4-hydroxylases, the key enzymes of collagen biosynthesis. Matrix Biol. 22(1): 15.

Myllyharju J. (2013) Prolyl 4-hydroxylases, master regulators of the hypoxia response. Acta Physiologica 208(2): 148-165.

Myllyharju J. (2008) Prolyl 4-hydroxylases, key enzymes in the synthesis of collagens and regulation of the response to hypoxia, and their roles as treatment targets. Ann Med 40(6): 402-417.

Myllyharju J & Kivirikko KI. (1997) Characterization of the iron- and 2-oxoglutarate-binding sites of human prolyl 4-hydroxylase. EMBO J 16(6): 1173-1180.

Myllyharju J & Kivirikko KI. (2001) Collagens and collagen-related diseases. Ann Med 33(1): 7-21.

Myllyharju J & Kivirikko KI. (2004) Collagens, modifying enzymes and their mutations in humans, flies, and worms. Trends Genet 20(1): 33-43.

Myllyharju J & Karppinen P. (2013) Hypoxia-inducible factor prolyl 4-hydroxylases: common and specific roles . 394(4): 435-448.

Myllyla R, Wang C, Heikkinen J, Juffer A, Lampela O, Risteli M, Ruotsalainen H, Salo A & Sipila L. (2007a) Expanding the lysyl hydroxylase toolbox: new insights into the localization and activities of lysyl hydroxylase 3 (LH3). J Cell Physiol 212(2): 323-329.

Myllyla R, Wang C, Heikkinen J, Juffer A, Lampela O, Risteli M, Ruotsalainen H, Salo A & Sipila L. (2007b) Expanding the lysyl hydroxylase toolbox: new insights into the localization and activities of lysyl hydroxylase 3 (LH3). J Cell Physiol 212(2): 323-329.

Ohh M, Takagi Y, Aso T, Stebbins CE, Pavletich NP, Zbar B, Conaway RC, Conaway JW & Kaelin WG, J. (1999) Synthetic peptides define critical contacts between elongin C, elongin B, and the von Hippel-Lindau protein. J Clin Invest 104(11): 1583-1591.

Onate SA, Boonyaratanakornkit V, Spencer TE, Tsai SY, Tsai MJ, Edwards DP & O'Malley BW. (1998) The steroid receptor coactivator-1 contains multiple receptors interacting and activation domains that cooperatively enhance the activation function 1 (AF1) and AF2 domains of steroid receptors. J Biol Chem 273(20): 12101-12108.

Ortega N & Werb Z. (2002) New functional roles for non-collagenous domains of basement membrane collagens. J Cell Sci 115(Pt 22): 4201-4214.

O'Shaughnessy L & Doyle S. (2011) Purification of proteins from baculovirus-infected insect cells. Methods Mol Biol 681: 295-309.

Pau MYM, Lipscomb JD & Solomon EI. (2007) Substrate activation for O2 reactions by oxidized metal centers in biology. Proc Natl Acad Sci U S A 104(47): 18355-18362.

Pekkala M, Hieta R, Bergmann U, Kivirikko KI, Wierenga RK & Myllyharju J. (2004) The peptide-substrate-binding domain of collagen prolyl 4-hydroxylases is a tetratricopeptide repeat domain with functional aromatic residues. J Biol Chem 279(50): 52255-52261.

Pektas S & Knapp MJ. (2013) Substrate preference of the HIF-prolyl hydroxylase-2 (PHD2) and substrate-induced conformational change. J Inorg Biochem 126: 55-60.

Pellequer J, Wager-Smith KA, Kay SA & Getzoff ED. (1998) Photoactive yellow protein: A structural prototype for the three-dimensional fold of the PAS domain superfamily. Proceedings of the National Academy of Sciences of the United States of America 95(11): 5884-5890.

Petoukhov MV, Franke D, Shkumatov AV, Tria G, Kikhney AG, Gajda M, Gorba C, Mertens HD, Konarev PV & Svergun DI. (2012) New developments in the ATSAS program package for small-angle scattering data analysis. J Appl Crystallogr 45(Pt 2): 342-350.

Petoukhov MV & Svergun DI. (2005) Global Rigid Body Modeling of Macromolecular Complexes against Small-Angle Scattering Data. Biophysical Journal 89(2): 1237-1250.

Pihlajaniemi T, Helaakoski T, Tasanen K, Myllyla R, Huhtala ML, Koivu J & Kivirikko KI. (1987) Molecular cloning of the beta-subunit of human prolyl 4-hydroxylase. This subunit and protein disulphide isomerase are products of the same gene. EMBO J 6(3): 643-649.

Pirneskoski A, Klappa P, Lobell M, Williamson RA, Byrne L, Alanen HI, Salo KE, Kivirikko KI, Freedman RB & Ruddock LW. (2004) Molecular characterization of the principal substrate binding site of the ubiquitous folding catalyst protein disulfide isomerase. J Biol Chem 279(11): 10374-10381.

Price JC, Barr EW, Hoffart LM, Krebs C & Bollinger JM,Jr. (2005) Kinetic dissection of the catalytic mechanism of taurine:alpha-ketoglutarate dioxygenase (TauD) from Escherichia coli. Biochemistry 44(22): 8138-8147.

Pronk S, Lindahl E & Kasson PM. (2014) Dynamic heterogeneity controls diffusion and viscosity near biological interfaces. Nature Communications 5: 3034.

Rasmussen JR. (1992) Effect of glycosylation on protein function. Curr Opin Struct Biol 2(5): 682-686.

Ratcliffe PJ. (2007) HIF-1 and HIF-2: working alone or together in hypoxia? J Clin Invest 117(4): 862-865.

Rosen MD, Venkatesan H, Peltier HM, Bembenek SD, Kanelakis KC, Zhao LX, Leonard BE, Hocutt FM, Wu X, Palomino HL, Brondstetter TI, Haugh PV, Cagnon L, Yan W, Liotta LA, Young A, Mirzadegan T, Shankley NP, Barrett TD & Rabinowitz MH. (2010) Benzimidazole-2-pyrazole HIF Prolyl 4-Hydroxylase Inhibitors as Oral Erythropoietin Secretagogues. ACS Med Chem Lett 1(9): 526-529.

Ruddock LW, Darby NJ, Klappa P & Freedman RB. (1998) The b′ domain provides the principal peptide-binding site of protein disulfide isomerase, but all domains contribute to binding of misfolded proteins — the EMBO Journal 17(4): 927-935.

Rupp. B. (2010) Biomolecular crystallography: principles, practice, and application to structural biology. Choice Reviews Online 47(10): 5661.

Ryland AL. (1958) X-ray diffraction. J Chem Educ 35(2): 80.

Salceda S & Caro J. (1997) Hypoxia-inducible factor 1alpha (HIF-1alpha) protein is rapidly degraded by the ubiquitin-proteasome system under normoxic conditions. Its stabilization by hypoxia depends on redox-induced changes. J Biol Chem 272(36): 22642-22647.

Schofield CJ & Ratcliffe PJ. (2004) Oxygen sensing by HIF hydroxylases. Nat Rev Mol Cell Biol 5(5): 343-354.

Schofield CJ & Zhang Z. (1999) Structural and mechanistic studies on 2-oxoglutarate-dependent oxygenases and related enzymes. Curr Opin Struct Biol 9(6): 722-731.

Semenza GL. (2001) HIF-1 and mechanisms of hypoxia sensing. Curr Opin Cell Biol. 13: 167-171.

Semenza GL. (2000) HIF-1: mediator of physiological and pathophysiological responses to hypoxia. J Appl Physiol (1985) 88(4): 1474-1480.

Semenza GL & Wang GL. (1992) A nuclear factor induced by hypoxia via de novo protein synthesis binds to the human erythropoietin gene enhancer at a site required for transcriptional activation. Mol Cell Biol 12(12): 5447-5454.

Shan X & Que L. (2005) Intermediates in the oxygenation of a nonheme diiron(II) complex, including the first evidence for a bound superoxo species. Proc Natl Acad Sci U S A 102(15): 5340.

Shi X & Jarvis DL. (2007) Protein N-glycosylation in the baculovirus-insect cell system. Curr Drug Targets 8(10): 1116-1125.

Shifman JM, Choi MH, Mihalas S, Mayo SL & Kennedy MB. (2006) Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) is activated by calmodulin with two bound calciums. Proc Natl Acad Sci USA 103(38): 13968.

Skou S, Gillilan RE & Ando N. (2014) Synchrotron-based small-angle X-ray scattering of proteins in solution. Nature protocols 9(7): 1727-1739.

Sosa Ldel V, Alfaro E, Santiago J, Narvaez D, Rosado MC, Rodriguez A, Gomez AM, Schreiter ER & Pastrana-Rios B. (2011) The structure, molecular dynamics, and energetics of centrin-melittin complex. Proteins 79(11): 3132-3143.

Stetefeld J, McKenna SA & Patel TR. (2016) Dynamic light scattering: a practical guide and applications in biomedical sciences. Biophysical reviews 8(4): 409-427.

Svergun DI. (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophys J 76(6): 2879-2886.

Svergun D, Barberato C & Koch MHJ. (1995) CRYSOL– a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. Journal of Applied Crystallography 28(6): 768-773.

Svergun D. (2013) Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules. Oxford: Oxford University Press.

Svergun DI & Koch MHJ. (2003) Small-angle scattering studies of biological macromolecules in solution. Reports on Progress in Physics 66(10): 1735-1782.

Tarhonskaya H, Chowdhury R, Leung IK, Loik ND, McCullagh JS, Claridge TD, Schofield CJ & Flashman E. (2014a) Investigating the contribution of the active site environment to the slow reaction of hypoxia-inducible factor prolyl hydroxylase domain 2 with oxygen. Biochem J 463(3): 363-372.

Tarhonskaya H, Szöllössi A, Leung IKH, Bush JT, Henry L, Chowdhury R, Iqbal A, Claridge TDW, Schofield CJ & Flashman E. (2014b) Studies on Deacetoxycephalosporin C Synthase Support a Consensus Mechanism for 2-Oxoglutarate Dependent Oxygenases. Biochemistry 53(15): 2483-2493.

Tasanen K, Parkkonen T, Chow LT, Kivirikko KI & Pihlajaniemi T. (1988) Characterization of the human gene for a polypeptide that acts both as the beta subunit of prolyl 4-hydroxylase and as a protein disulfide isomerase. J Biol Chem 263(31): 16218-16224.

Taylor GL. (2010) Introduction to phasing. Acta crystallographica.Section D, Biological crystallography 66: 325-338.

Taylor MS. (2001) Characterization and comparative analysis of the EGLN gene family. Gene 275(1): 125-132.

Tian H, McKnight SL & Russell DW. (1997) Endothelial PAS domain protein 1 (EPAS1), a transcription factor selectively expressed in endothelial cells. Genes Dev. 11(1): 82.

Trastoy B, Klontz E, Orwenyo J, Marina A, Wang L, Sundberg EJ & Guerin ME. (2018) Structural basis for the recognition of complex-type N-glycans by Endoglycosidase S. Nature Communications 9(1): 1874.

Trewick SC, Henshaw TF, Hausinger RP, Lindahl T & Sedgwick B. (2002) Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage. Nature 419(6903): 174-178.

Tusnady GE & Simon I. (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283(2): 489-506.

Vaillancourt FH, Yeh E, Vosburg DA, O'Connor SE & Walsh CT. (2005) Cryptic chlorination by a non-haem iron enzyme during cyclopropyl amino acid biosynthesis. Nature 436(7054): 1191-1194.

Vallee BL & Williams RJ. (1968) Metalloenzymes: the entatic nature of their active sites. Proc Natl Acad Sci U S A 59(2): 498-505.

Varshavsky A. (2011) The N-end rule pathway and regulation by proteolysis. Protein Sci 20(8): 1298-1345.

Velez Perez JA, Guzman O & Navarro-Garcia F. (2013) Steric contribution of macromolecular crowding to the time and activation energy for preprotein translocation

across the endoplasmic reticulum membrane. Phys Rev E Stat Nonlin Soft Matter Phys 88(1): 012725.

Vuori K, Pihlajaniemi T, Myllyla R & Kivirikko KI. (1992) Site-directed mutagenesis of human protein disulphide isomerase: effect on the assembly, activity and endoplasmic reticulum retention of human prolyl 4-hydroxylase in Spodoptera frugiperda insect cells. EMBO J 11(11): 4213-4217.

Wallis AK, Sidhu A, Byrne LJ, Howard MJ, Ruddock LW, Williamson RA & Freedman RB. (2009) The ligand-binding b' domain of human protein disulphide-isomerase mediates homodimerization. Protein science : a publication of the Protein Society 18(12): 2569-2577.

Wang C, Yu J, Huo L, Wang L, Feng W & Wang CC. (2012) Human protein-disulfide isomerase is a redox-regulated chaperone activated by oxidation of domain a'. J Biol Chem 287(2): 1139-1149.

Wang GL, Jiang BH, Rue EA & Semenza GL. (1995) Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension. Proc Natl Acad Sci U S A. 92(12): 5510–5514.

Wang X & Schwarz TL. (2009) The mechanism of Ca2+ -dependent regulation of kinesin-mediated mitochondrial motility. Cell 136(1): 163-174.

Weller MT, Overton T, Rourke J & Armstrong FA. (2014) Inorganic Chemistry. Oxford: Oxford University Press.

Wolfenden R. (2014) Massive thermal acceleration of the emergence of primordial chemistry, the incidence of spontaneous mutation, and the evolution of enzymes. The Journal of biological chemistry 289(44): 30198-30204.

Xie L, Xiao K, Whalen EJ, Forrester MT, Freeman RS, Fong G, Gygi SP, Lefkowitz RJ & Stamler JS. (2009) Oxygen-regulated beta(2)-adrenergic receptor hydroxylation by EGLN3 and ubiquitylation by pVHL. Sci Signal 2(78): ra33.

Zhang M, Tanaka T & Ikura M. (1995) Calcium-induced conformational transition revealed by the solution structure of apo calmodulin. Nat Struct Biol 2(9): 758-767.

Zhang Z, Ren J, Clifton IJ & Schofield CJ. (2004) Crystal Structure and Mechanistic Implications of 1-Aminocyclopropane-1-Carboxylic Acid Oxidase—The Ethylene-Forming Enzyme. Chem Biol 11(10): 1383-1394.